The American Economic Review

ARTICLES

WILLIAM VICKREY — Today's Task for Economists

GIUSEPPE BERTOLA AND ALLAN DRAZEN
Trigger Points and Budget Cuts: Explaining the Effects of Fiscal Austerity

JOSEPH E. HARRINGTON, JR.
Economic Policy, Economic Performance, and Elections

A. ANDREW JOHN, ROWENA A. PECCHENINO, AND STACEY L. SCHREFT
The Macroeconomics of Dr. Strangelove

NOBUHIRO KIYOTAKI AND RANDALL WRIGHT
A Search-Theoretic Approach to Monetary Economics

ANIL K KASHYAP, JEREMY C. STEIN, AND DAVID W. WILCOX
Monetary Policy and Credit Conditions: Evidence from the Composition of External Finance

MICHELLE R. GARFINKEL AND SEONGHWAN OH
Strategic Discipline in Monetary Policy with Private Information: Optimal Targeting Horizons

GREGORY K. DOW — Why Capital Hires Labor: A Bargaining Perspective

IAN L. GALE AND THOMAS J. HOLMES
Advance-Purchase Discounts and Monopoly Allocation of Capacity

KALA KRISHNA — Auctions with Endogenous Valuations: The Persistence of Monopoly Revisited

RICHARD ARNOTT, ANDRÉ DE PALMA, AND ROBIN LINDSEY
A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand

RICHARD H. CLARIDA
Entry, Dumping, and Shakeout

STEPHEN G. BRONARS AND DONALD R. DEERE
Union Organizing Activity, Firm Growth, and the Business Cycle

MICHAEL R RANSOM
Seniority and Monopsony in the Academic Labor Market

SHORTER PAPERS: R. A. Margo and T. A. Finegan; M. Honig and R. K. Filer; E. J. Dockner and G. Feichtinger; G. D. Rudebusch; R. S. Pindyck; H. Bester; M. R. Baye, D. Kovenock, and C. G. de Vries; X. Yang and B. J. Heijdra; A. K. Dixit and J. E. Stiglitz; J. Agell and B.-C. Ysander; S. Fischer and L. Summers.

MARCH 1993

THE AMERICAN ECONOMIC ASSOCIATION

•Printed at Banta Company, Menasha, Wisconsin.

•Copyright © American Economic Association 1993. All rights reserved.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002–8282), March 1993, Vol. 83, No. 1, is published six times a year (March, May, June, September, December semimonthly) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: $47.40, $56.90, or $66.40 depending on income. A membership also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add $23.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 2014 Broadway, Suite 305, Nashville, TN 37203.

Founded in 1885

Officers

President
ZVI GRILICHES
Harvard University

President-elect
AMARTYA SEN
Harvard University

Vice-Presidents
FRANCINE BLAU
University of Illinois
LESTER C. THUROW
Massachusetts Institute of Technology

Secretary-Treasurer
C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review
ORLEY C. ASHENFELTER
Princeton University

Editor of The Journal of Economic Literature
JOHN PENCAVEL
Stanford University

Editor of The Journal of Economic Perspectives
JOSEPH E. STIGLITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee
MICHAEL J. PIORE
Massachusetts Institute of Technology
GAVIN WRIGHT
Stanford University
ALAN J. AUERBACH
University of Pennsylvania
JOHN B. TAYLOR
Stanford University
WILLIAM A. DARITY, JR.
University of North Carolina
RICHARD L. SCHMALENSEE
Massachusetts Institute of Technology

EX OFFICIO Members
THOMAS C. SCHELLING
University of Maryland
WILLIAM VICKREY
Columbia University

THE AMERICAN ECONOMIC REVIEW

Editor
ORLEY ASHENFELTER

Co-Editors
JOHN Y. CAMPBELL
ROGER H. GORDON
PAUL R. MILGROM

Production Editor
J. DAVID BALDWIN

Board of Editors

JOSEPH G. ALTONJI
ALAN J. AUERBACH
KYLE W. BAGWELL
ROBIN W. BOADWAY
TIMOTHY F. BRESNAHAN
LORNE H. CARMICHAEL
STEPHEN G. CECCHETTI
LAWRENCE J. CHRISTIANO
GEORGE W. EVANS
MARJORIE A. FLAVIN
NANCY GALLINI
JO ANNA GRAY
DANIEL S. HAMERMESH
ROBERT H. HAVEMAN
ROBERT J. HODRICK
KEVIN D. HOOVER
R. MARK ISAAC
JOHN H. KAGEL
JOHN F. KENNAN
R. PRESTON McAFEE
JOHN McMILLAN
JOHN ROBERTS
RICHARD ROLL
THOMAS ROMER
SUZANNE A. SCOTCHMER
ROBERT H. TOPEL
HAL R. VARIAN
KENNETH WEST
JOHN D. WILSON
LESLIE YOUNG

March 1993

VOLUME 83, NUMBER 1

Articles

Today's Task for Economists *William Vickrey* 1

Trigger Points and Budget Cuts: Explaining the Effects of Fiscal Austerity *Giuseppe Bertola and Allan Drazen* 11

Economic Policy, Economic Performance, and Elections *Joseph E. Harrington, Jr.* 27

The Macroeconomics of Dr. Strangelove *A. Andrew John, Rowena A. Pecchenino, and Stacey L. Schreft* 43

A Search-Theoretic Approach to Monetary Economics *Nobuhiro Kiyotaki and Randall Wright* 63

Monetary Policy and Credit Conditions: Evidence from the Composition of External Finance *Anil K Kashyap, Jeremy C. Stein, and David W. Wilcox* 78

Strategic Discipline in Monetary Policy with Private Information: Optimal Targeting Horizons *Michelle R. Garfinkel and Seonghwan Oh* 99

Why Capital Hires Labor: A Bargaining Perspective *Gregory K. Dow* 118

Advance-Purchase Discounts and Monopoly Allocation of Capacity *Ian L. Gale and Thomas J. Holmes* 135

Auctions with Endogenous Valuations: The Persistence of Monopoly Revisited *Kala Krishna* 147

A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand *Richard J. Arnott, Andre de Palma, and Robin Lindsey* 161

Entry, Dumping, and Shakeout *Richard H. Clarida* 180

Union Organizing Activity, Firm Growth, and the Business Cycle *Stephen G. Bronars and Donald R. Deere* 203

Seniority and Monopsony in the Academic Labor Market *Michael R Ransom* 221

Shorter Papers

The Decline in Black Teenage Labor-Force Participation in the South, 1900–1970: The Role of Schooling — *Robert A. Margo and T. Aldrich Finegan* 234
Causes of Intercity Variation in Homelessness — *Marjorie Honig and Randall K. Filer* 248
Cyclical Consumption Patterns and Rational Addiction — *Engelbert J. Dockner and Gustav Feichtinger* 256
The Uncertain Unit Root in Real GNP — *Glenn D. Rudebusch* 264
A Note on Competitive Investment under Uncertainty — *Robert S. Pindyck* 273
Bargaining versus Price Competition in Markets with Quality Uncertainty — *Helmut Bester* 278
Rigging the Lobbying Process: An Application of the All-Pay Auction — *Michael R. Baye, Dan Kovenock, and Casper G. de Vries* 289
Monopolistic Competition and Optimum Product Diversity
Comment — *Xiaokai Yang and Ben J. Heijdra* 295
Reply — *Avinash K. Dixit and Joseph E. Stiglitz* 302
Should Governments Learn to Live with Inflation?
Comment — *Jonas Agell and Bengt-Christer Ysander* 305
Reply — *Stanley Fischer and Lawrence Summers* 312

Erratum

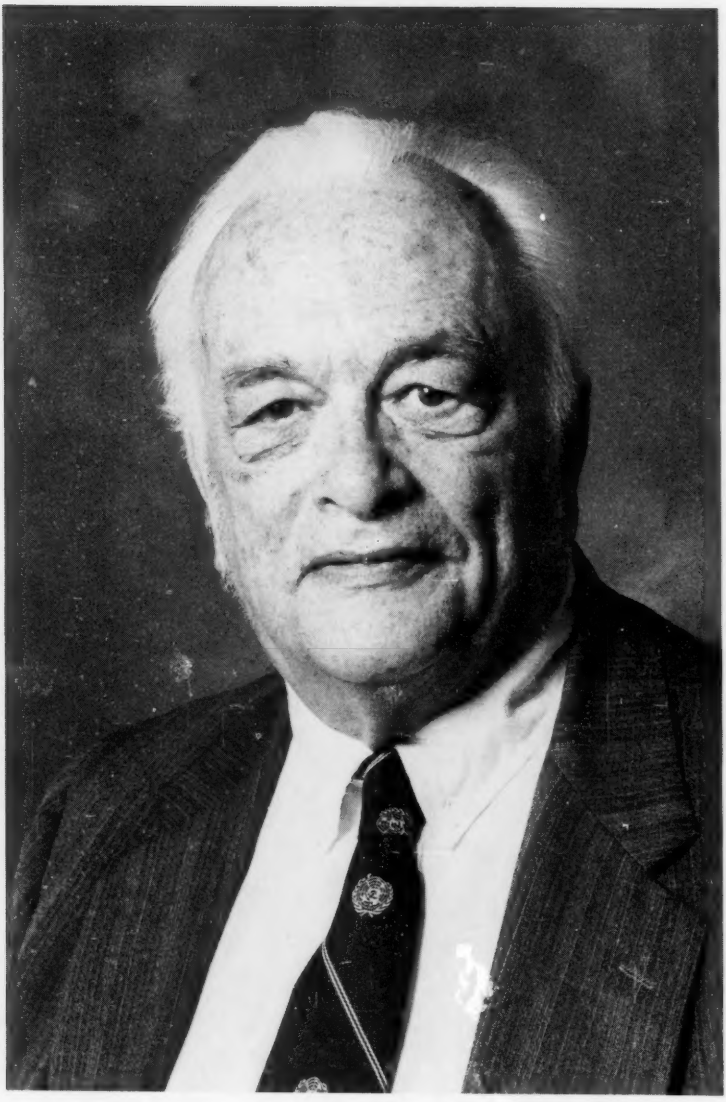
Government Size and Economic Growth: A New Framework and Some Evidence from Cross-Section and Time-Series Data: Reply — *Rati Ram* 314

The following Statement of Ownership, Management, and Circulation is provided in accordance with the requirements, as contained in 39 U.S. Code 3685. The *American Economic Review* is owned, managed, and published by the American Economic Association, a nonprofit scientific organization, located at 2014 Broadway, Suite 305, Nashville, Davidson County, Tennessee 37203–2418. The Editor is Professor Orley Ashenfelter, *American Economic Review*, 209 Nassau Street, Princeton, NJ 08542–4607. During the preceding 12 months the average number of copies printed for each issue was 27,900; the average paid circulation, 25,539; the average free distribution, 123; the average number of copies distributed, 25,662. Corresponding figures for the issue published nearest to the filing date: total number of copies printed, 27,500; total paid circulation, 25,644; total free distribution, 119; total distribution, 25,763.

- Submit manuscripts (4 copies), single-sided, double-spaced, to: Orley Ashenfelter, Editor, *AER*, 209 Nassau Street, Princeton, NJ 08542–4607.
- Authorship should be identified only on a removable cover page; the anonymous text should begin on the following page.
- Submission fee: $50 for members; $100 for nonmembers. Please pay with a check or money order payable in United States dollars. Foreign (including Canadian) payments must be in the form of a check drawn on a United States bank payable in United States dollars.
- Style guides will be provided upon request.

It is the policy of the *American Economic Review* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Details of the computations sufficient to permit replication must be provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary, or if, for some other reason, the above requirements cannot be met.

Number 94 of a series of photographs of past presidents of the Association



William Fulbright

Today's Task for Economists[†]

By WILLIAM VICKREY*

I. The Allocation Task of Yesteryear

Nearly two score years ago, on the occasion of Columbia's bicentennial celebration, Sir Dennis Robertson gave an address entitled "What Do Economists Economize," the burden of which was that, since presumably economists are the most expert economizers, they should economize the most precious thing in the world, namely, love, or altruism. This would be done in part by so arranging things that in the ordinary conduct of life individual choices made on the basis of self-interest in terms of market prices would at least be consistent with maximizing social welfare, so that the exercise of scarce resources of altruism could be concentrated on situations where Adam Smith's unseen hand could not be made to serve. To me, one implication of this was that economists should see to it that market prices correctly reflect the relevant marginal social cost of various alternatives. I have devoted a major part of my career to the promotion of such marginal-cost pricing, but thus far with a notable lack of practical success outside academia.

At the time of Robertson's address, indeed, there was a certain euphoria prevailing among at least part of the economics profession over the prospect of curbing the business cycle and maintaining a high level of economic activity through Keynesian fiscal policy. Under these circumstances it was reasonable to think that the chief remaining job of the economist was to assure a Pareto-efficient allocation of a given aggregate of resources. The event, however, proved otherwise. The conventional wisdom of regarding budget deficits as improvident prodigality, and government debt as the legacy of a craven deferral of burden to the future, resumed command.

One eminent economist is said to have remarked, in effect, that it was the function of the science of public finance to see to it that nothing of importance is ever done or left undone merely for financial reasons. Alas, the financial reasons have thus far carried the day, and we have not had anything approaching real full employment since the Korean War, or indeed in peacetime at any time since 1925, if then, at least in terms of the Beveridge definition of full employment as a situation wherein there are at least as many unfilled job openings as there are unemployed individuals seeking work.

In the Eisenhower years, the conventional wisdom held sway in spite of the absence of serious contraindications to the Keynesian prescription. In the 1960's, the simple Keynesian analysis began to be called into question by the emergence of stagflation, a phenomenon not contemplated by the earlier Keynesian models. A new relationship, the Phillips curve, relating the evolution of inflation to the level of unemployment was added to the economists' armamentarium, with its "non-inflation-accelerating rate of unemployment" or NIARU.

This NIARU is of course not a fixed datum, but varies over time and place according to the sociopolitical ambience, the mechanics of the labor market, and the vigor of competition. It may have been rising over time as a result of the increased sophistication and differentiation of products, real and factitious, giving sellers, as the ones most knowledgeable about the characteristics of their products and their markets, considerable leeway to raise their prices without unacceptable loss of sales. This process is ultimately held in check only

[†]Presidential address prepared for the one-hundred fifth meeting of the American Economic Association, January 6, 1993, Anaheim, CA.

*Professor of Economics Emeritus, Columbia University, New York, NY 10027.

by the presence of underutilized labor and other resources. Currently in the United States the NIARU appears to be around 4–6 percent.

In some quarters this NIARU has even been termed the "natural" rate of unemployment, in one of the most vicious euphemisms ever coined. Some have even gone so far as to define "full employment" as being the NIARU. But while 5-percent unemployment might be barely tolerable if it meant that everyone would be taking an additional two weeks of vacation every year without pay, it is totally unacceptable as a social goal when it means unemployment rates of 10, 20, or even 40 percent among disadvantaged groups, with resulting increases in poverty, homelessness, poor health, drug addiction, and crime. Yet the hard political fact is that at such a NIARU the great majority of the voting population, including most of the politically active upper and middle classes, will have relatively little personal experience of severe unemployment, while nearly everyone will have some direct experience of inflation. Many seem to feel that if only prices would stop rising they would benefit correspondingly by having their income go further, giving relatively little thought to the effect on their incomes. Even those with large mortgages or other debts, who would actually gain from inflation, tend to concur in the notion that they suffer from it. It is thus extremely difficult to get political support for anti-unemployment measures that are perceived as involving a threat of inflation, at least until unemployment reaches 7 percent or more, at which point unemployment becomes a more widespread threat.

Actually it is the uncertainty as to the rate of inflation, and not its level, that does the damage. An assured, moderate rate of inflation can be adapted to by adjusting nominal rates of interest and the terms of long-term contracts involving money payments. The "menu cost" of changing price tags and catalog quotations is probably less important than the mental effort required of consumers in forming an idea of what an appropriate current price is for infrequently purchased items, such as furniture or clothing. An inflation rate assured to stay between 5 percent and 6 percent, say, might even have advantages. Monetary policy would be more powerful in stemming a downturn in that very low and even negative real rates of interest would become feasible as a stimulus to investment. It might in principle be easier to keep inflation within a 1-percent range between 5 percent and 6 percent, than to keep it within a 2-percent range between -1 percent and $+1$ percent, given the smaller real value of non-interest-bearing moneys in circulation, even allowing for the superior political focusing power of a target of 0 percent as compared to one of 5.5 percent.

The base of the income tax would be broadened also, making it possible to have a tax that is more progressive and more productive of revenue with lower marginal rates and less of a distortionary effect. A tax based on nominal accrued income would in effect be a tax on a base consisting of real income plus a percentage of net worth. While this is not what is meant by an ideologically pure income tax, in terms of its practical effects it can be deemed a superior tax.

It is the possibility of substantial changes in the rate of inflation, either up or down, that does the damage. Such changes involve a disappointment of expectations and a redistribution of wealth and income derived from a given national product that is capricious and often inequitable, but it does not of itself substantially reduce the amount to be distributed. Unemployment, on the other hand, directly and definitely reduces the total product to be distributed. Unanticipated changes in the rate of inflation, up or down, may be considered to be a form of legitimized embezzlement, whereas unemployment is vandalism.

Nevertheless, the stance of the politico-financial establishment is still to look at the bottom line as the ultimate reality, whether of the corporation or the national budget, and since money is the measure of all good and evil in this kind of calculus, anything that impugns the value of money is viewed as a kind of sacrilege reinforced by a lurking fear of starting down a slippery slope to

hyperinflation. We find the Federal Reserve System poised to slam on the brakes at the first sign of a resurgence of inflation, a posture not calculated to inspire investment in durable capital.

On the political side, we see the House voting by a substantial majority in favor of a constitutional amendment to require a balanced budget, fortunately falling short of the required two-thirds. This was done in spite of the fact that the nominal budget as currently computed is not a valid measure of any significant economic quantity. The nominal deficit would be reduced by selling the Pentagon to a life insurance company subject to a long-term lease-back and repurchase option; this at least would do no harm, unlike the sale of natural resources to private exploiters which would actually decrease the real heritage handed down to the future, on the pretext of reducing the transfer requirements embodied in the national debt.

II. Recycling Savings Through Public Capital Formation

From a classical standpoint, of course, the difficulty is that no account is taken of the distinction between transactions on current account and on capital account. If AT&T, General Motors, and households had been constrained to operate under the restrictions of the proposed balanced-budget amendment, we would now have far fewer telephones, automobiles, and houses. A capital budget, with a vast expansion of government capital outlays on roads, bridges, research, education, and the like, financed by borrowing, might go a considerable way toward improving the unemployment situation. But there is no assurance that it could do the whole job.

III. Eliminating the Corporation Income Tax

Other classical approaches to improving the unemployment situation exist but have their own political opposition and in any case are too weak to make much of a dent in a very large need. One such measure would be the abolition of the corporation income tax, which is by far the most serious hurdle in the way of private capital formation of a kind requiring equity funding. Unlike the capital-gains tax, the corporate income tax is a tax largely above or before the market, requiring a rate of return on investment sufficient to cover the corporation tax and leave a rate of return after tax comparable to other investments, whereas the capital-gains tax operates largely as a reduction in the return to the investor after or below the market, comparable to the reduction of net income to the taxpayer resulting from the personal income tax on other income. In addition, the corporation tax causes inefficient allocation of investment between equity-type and loan-type investments; it encourages thin equity and resulting bankruptcies and reorganizations, and it lubricates takeovers and mergers of dubious intrinsic merit.

Reduction of the tax on capital gains, on the other hand, might actually depress economic activity if the additional savings out of the tax reduction were to exceed the additional capital formation induced. This is the more likely in that most of the tax reduction is likely to be saved immediately, whereas the inducement to capital formation is in terms of a tax reduction in a relatively remote future, subject to legislative vicissitudes. At best, special treatment of capital gains greatly increases the complexity of the tax law and diverts investment flows from their most efficient use. There is nothing to indicate that investments likely to yield returns in forms defined by the tax code as capital gains will have any superior social value: gains from land speculation, in particular, add nothing to the real availability of resources.

As for the corporate income tax, in spite of its many defects from the standpoint of economic efficiency, it has enormous political popularity due to the fact that nearly everyone thinks that it is paid by someone else. Indeed economists have differed widely in their assignment of the "burden" of the tax, owing to a failure to specify, or even to consider, the macroeconomic policy changes necessarily involved in a change in the tax. Unlike most other taxes, the corporation tax

inflicts a double whammy on the economy in that it both extracts income from the stream of purchasing power and reduces the recycling of savings through investment. If imposed on a revenue-neutral basis it causes unemployment, while if a budgetary adjustment is made to maintain employment constant, its burden can be thought of as falling on future wage earners, who will have less capital with which to work.

Problems of the deferral of income through undistributed profits, as well as the deferral of taxation to the time of realization of capital gains, would ideally be met by putting the personal income tax on a cumulative basis, along lines I developed while working with Carl Shoup in 1938, whereby the deferral of the reporting of income, by whatever means, merely involves the borrowing of the deferred tax at a suitable rate of interest. About two-thirds of the internal revenue code would become redundant, with the possible exception of the need to deal with the international jet set and revolving-door marriages; large numbers of tax techies would be able to apply themselves to more productive employment.

Failing this, an approximation to a level playing field might be had by imposing a small annual tax on the accumulated undistributed surplus of corporations, roughly equal to the interest on the stockholders' postponed individual income tax. Similarly, there should be a surcharge on realized capital gains, proportionate to the length of time held, to offset the gain from the deferral of the tax.

If there is nevertheless a need to cater to a political demand for something that can be labeled a corporation tax, this might be satisfied by levying a corporation tax on dividends, interest, and retained earnings at a rate corresponding to the first-bracket rate of the individual income tax and exempting such interest and dividends from this "normal" rate, going back to the pre-1934 practice of dividing the income tax into a normal tax and a progressive surtax. To even things up neatly, normal tax paid on other forms of income should be deductible in computing the base for the progressive surtax paid by a minority of taxpayers. It would still be appropriate to have an undistributed surplus tax to correspond to this surtax.

IV. Tax-Exempt Bonds

Another measure that might slightly improve investment allocation would be to replace the exemption of interest on state and local bonds by a taxable tax credit at a rate that would maintain the market value of the bonds. Low-bracket taxpayers would be little affected, while the entire loss of revenue to the Treasury would accrue as a subsidy to the issuers. Upper-bracket taxpayers would no longer have an incentive to invest in such bonds rather than in riskier investments more suitable to their status.

V. Taxing Imputed Income

A more important but politically more difficult measure would be to require the inclusion in taxable income of the rental value of owner-occupied residences. This would not only improve the equity and progressivity of the income tax but go a substantial way toward making more units available for rental and, to a modest extent, promoting the construction of additional affordable rental housing and abating the problem of homelessness. A similar case can be made for including in the income tax base a net rental value of nonbusiness automobiles (equal to interest on the market value of the car), in this case reducing the discrimination against the use of public transit.

VI. Shifting Property Taxes from Improvements to Land

A measure that could provide a powerful stimulus to investment in property improvements would be to replace part or all of the property tax by a tax on land value only, a proposal that can be traced all the way back to François Quesnay and the French physiocrats but which is more recently associated with the name of Henry George. This

would remove the very serious deterrent effect of the property tax on improvements. Unfortunately from the standpoint of a national employment policy, this tax is largely levied by local governments, which are often constrained by constitutional provisions or state laws. Nevertheless some means of bringing pressure to bear on these governments to make this change might be found. Some Pennsylvania governments are already doing this. When levied for municipal purposes, it might be appropriate to exempt from the tax a flat amount per square foot as representing the value of circumambient agricultural land for which the urban government can claim no credit; this would also mitigate discriminations at jurisdictional boundaries.

It is perhaps worth noting that the significance of a government debt would be drastically different in a community relying exclusively on a land-value tax. Such a debt would in effect be a collective mortgage on the land, especially if it can be assumed that land values in the community will vary proportionately over time. Since the interest on the community debt will generally be lower than interest charged on individual mortgages, it can be in the general interest of all the taxpayers of the community for the government to borrow as much as the market will take, even to finance current outlays, provided a suitable margin is left to deal with emergencies. On the other hand such debt financing performs no recycling of savings, there is no room for Keynesian fiscal policy, and Ricardian equivalence is in full sway. This does not detract, however, from the powerful stimulating effect of a reduction in the tax on improvements.

VII. Limitations of "Supply-Side" Measures

Under current conditions, however, such "supply-side" measures designed to operate by reducing the cost of capital are likely to be severely limited in their effect as long as nearly all types of capital facilities are idle or underutilized. Very little "widening" investment is likely to take place as long as there is excess capacity in place. At most, some "deepening" investment in new products or technologies may take place, or there may be corners of the economy where relatively rapid growth has kept capacity fully utilized. Even in such cases, investment in capital facilities may depend more on appraisals of an uncertain market for the product than on the cost of capital.

This is likely to be true not only of tax policy but even more of monetary policy. In any attempt to emerge from present rates of unemployment even only down to the NIARU within any reasonable time period, monetary policy is likely to prove a weak reed, sometimes aptly described as pushing on a string. The main difficulty is that monetary policy bears primarily on short-term interest rates and credit availability and in its usual practice does not directly control long-term rates, which are the important rates for most decisions involving real durable capital formation. The posture of the Federal Reserve System in holding itself ready to slam on the brakes at the first sign of resurgent inflation is poorly adapted to bringing long-term rates down. It does not appear that the Fed has either the will or the resources to do enough about long-term rates to do very much to increase capital formation, especially when idle and underused capacity pervades much of the economy.

VIII. Savings Recycling by Government

This brings us inevitably around to fiscal policy. Here it is necessary to stop thinking of the conventional nominal budget deficit, or even of a current-account deficit in a budget drawn up in terms of distinguishing capital and current-account items, and to start thinking of fiscal policy in terms of its role in recycling savings, in excess of what is recycled by private investment, into the stream of purchasing power. The conventional wisdom seems to argue that increased employment requires that the economy grow, growth requires investment, and investment requires savings; therefore let's encourage saving through IRA's, tax expenditure rather than income, and tighten our

belts to restore the economy to its normal state of health.

It doesn't work that way. Savings are not like a sack of potatoes which if not sold at the current price will stay on hand and put a downward pressure on the price until sold. Savings not immediately taken up to create capital simply vanish in reduced income, without even exerting a downward pressure on interest rates. If I yield to the allurements of tax concessions to IRA's to the point of not having my hair cut, this puts $8 more in my bank account, but $8 less in the barber's account; there is nothing that makes it any easier for anyone to obtain funds with which to create capital, nor anything that makes the prospect more attractive. As Gertrude Stein remarked, "the money is always there, it's the pockets that keep changing." If the barber reacts by curtaling his consumption, this further reduces national income and saving. I may succeed in my attempt to save, but only by reducing the saving of others by even more. Savings are an extremely perishable entity. Say's law fails as soon as part of the income generated in the process of producing the supply is shunted off into savings that fail to get converted into new capital goods.

On the other hand, if some genius invents a new product or process and obtains a credit or borrows the funds needed to finance the capital involved in its production, this added real wealth is, ipso facto, someone's saving. Instead of Say's law, we have "capital formation creates its own saving." Similarly, if the government borrows funds created by credit expansion and recycles them into purchasing power through outlays, whether on current or capital account, this creates both income out of which additional savings will be attempted and demand that may induce the private investment to meet it.

Not all deficit financing, however, results in recycling of savings, whether measured by the current capriciously defined nominal deficit or by a more rational definition involving accounting for government assets. We have seen that in a community relying exclusively on a land tax, recycling does not take place. Nor would the sale of the Pentagon, or the purchase of an office building currently being rented by the government, offset by bond transactions, involve any change in the level of recycling. Government recycling is in principle the excess of those government outlays that are regarded by their recipients as income over those government receipts that are regarded by their payors as reductions in their disposable income. Even this is subject to some caveats: if government investment in a power plant, for example, substitutes for investment that would otherwise have been made by private enterprise, there is no net recycling.

On the whole, however, recycling tends to vary in rough correlation with the nominal deficit, and the strength of the notion in the minds of the public and their representatives that deficits are bad and that the "budget" should be balanced may make it difficult to achieve an adequate level of recycling. Some help in this respect may be obtained by going to a capital budget system, in which balance would be sought only for the current-account part of the budget, borrowing for the capital account being justified by comparisons with corresponding private practices and by the thought that future generations being burdened with the debt would also reap benefits from the capital passed on to them. While this may constrain choice away from what rational voters would have chosen as the optimal level of government capital formation, there would seem to be sufficient scope for government capital investment to provide sufficient recycling to bring about full employment, particularly if investments in education, research, space exploration, and the like are considered eligible for treatment as capital investment. Some of these projects, even if they would not stand scrutiny aside from their function in justifying income recycling, may nevertheless have the same kind of justification as the building of the Egyptian pyramids had for Keynes. On general welfare grounds, one might well prefer recycling in terms of borrowing to finance health care to borrowing to finance space stations, but if borrowing for health care is deemed to create an ideologically sinful current-account deficit, space stations it will have to be.

IX. The Need for Direct Inflation Control

Long before the economy reaches a really satisfactory level of full employment, however, as employment gets to the NIARU level, and inflation threatens to accelerate, the Fed is likely to try to slam on the brakes, and demands for a more stringent budget balancing and cutback of "government waste" are likely to be heard in the halls of Congress. To get anywhere near a satisfactory level of unemployment, some method of dealing with inflation will have to be devised. We are short of tools.

In effect, the economy can be thought of as having three major parameters that we would like to control: the level of employment of human and other resources, the price level, and the division of the resulting total product between provision for current wants and investment in growth and the future. At the same time, we have only two major policy tools: monetary and fiscal policy. In an era when inflation was not a threat, one could think of these two tools as controlling the level of employment and the rate of growth, with low interest rates combined with a deficit or surplus sufficient to maintain full employment leading to high investment and growth, and conversely. However, with a need to control inflation as well, relying on only two dimensions of control is like trying to fly an airplane without ailerons, which were the third dimension of control that was the key to the success of the Wright brothers. A new tool is needed.

Over the past three decades a number of proposals for direct control of inflation have been made, but none has achieved general acceptance. Wartime control of specific prices, accompanied by rationing, was accepted as an emergency measure and worked in part because of patriotic willingness to conform and in part because, being temporary, past prices could be continued without becoming absurd. As a permanent scheme this is probably unworkable and certainly unacceptable. More recent schemes have involved tax incentives of various kinds to provide a countervailing downward pressure against the inherent inflationary tendency of an imperfectly competitive system. Such schemes have generally suffered from difficulties in measuring price changes at an individual-firm level, capriciousness of results when tied to such taxes as the corporation income tax, and possible time lags in adjusting the strength of the incentives to changing circumstances.

X. Market-Based Inflation-Control Plans

A few years ago David Colander came to visit me and reported on a proposal by Abba Lerner for a market in rights to raise prices. Those wishing to raise their prices would be required to purchase the right from those prepared to lower their prices, thus assuring a constant overall price level. While this neatly circumvents the problem of adjusting the strength of incentive to changing inflationary pressures, the problem remains of how to measure price changes in the face of quality changes, new products, and variations in the terms of sale such as delivery, reliability, service, credit terms, tie-in sales, and the like.

More pregnant was the question of how to deal with cases in which prices paid to suppliers have risen. A somewhat similar problem arises with gross receipts taxes, which discriminate in favor of vertically integrated operations and against situations in which the product passes through several hands on the way to the market. In Europe this problem has been solved by shifting from gross receipts taxes and retail sales taxes to value-added taxes, which immediately suggests that instead of a market in rights to raise prices we have a market in rights to value added.

XI. Control with Marketable Gross Markup Warrants

For semantic reasons I have chosen to speak in terms of "gross markups" rather than value added, as being more suggestive of something to be restrained rather than promoted. In principle, gross markups simply refers to the excess of sales revenue over amounts paid for nonprime inputs. In operation, warrants for gross markups for a

prospective accounting period would be issued to each firm on the basis of the gross markups for a corresponding preceding period, plus or minus adjustments for changes in prime inputs such as labor and invested capital. These warrants would be issued in sufficient total face value to correspond to the value at a desired price level of the output expected to be produced by the inputs against which the warrants were issued. They would be freely tradable for cash in a competitive market, and if at the end of the accounting period a firm is found to have retained or acquired fewer warrants than the actual amount of its gross markups for the period, a penalty tax would be assessed. This tax would not be a substantial source of revenue, but would serve merely as an enforcement device. It could be set at a level fairly certain to be higher than the market price of the warrants.

Adjustment of the warrant issue for changes in investment could be made simply on the basis of a uniform percentage of such change. Adjustment for changes in employment is somewhat more difficult: a flat amount per employee or man-hour takes too little account of variations in qualifications, while to allow adjustments equal to payrolls would run a danger of allowing inflationary wage increases. Some formula such as a percentage of payrolls plus a flat amount per employee might be satisfactory; such a formula would involve a certain bias in favor of the employment of low-skill labor, which may be considered desirable in view of the fact that this is where the unemployment problem is most serious.

Administration would seem to pose no insurmountable problems. Determination of gross markups is essentially no different than the assessment of a value-added tax such as is widespread in Europe. Adjustment for investment can be made on the basis of accounts already needed for income-tax purposes, while adjustments for employment can be related to the social-security records. Some special methods may have to be developed for dealing with the self-employed and very small firms, and possibly some classes of firms could be excluded from the scheme, as is sometimes done with the value-added tax.

XII. Prospects for Rapidly Reaching Genuine Full Employment

With such a scheme in place, what can we plan for in terms of getting from where we are to full employment? Currently unemployment is reported as about 7.5 percent, and full employment can be reckoned at about 1.5 percent, giving a slack to be made up of 6 percent. Using Okun's ratio of percentage change in GNP to percentage change in reported unemployment of 2.5, we have a slack of 15 percent to be made up. If this slack can be taken up within two years, this will be 7.5 percent per year; if to this we add 2.5 percent for growth in the labor force and in productivity, we get 10-percent annual growth in GNP over two years. After two years, we hit the full-employment ceiling, and growth thereafter will be limited to the labor force and productivity factor, possibly between 2 percent and 4 percent.

Is public finance up to the job of reaching the goals thus defined in terms of the limits of our real resources? Possibly, but it requires breaking new ground. One would have to begin with increasing government recycling as rapidly as possible by 8–10 percent of GNP in order to inaugurate the 10-percent growth rate. How rapidly this could be done would of course depend on the political and legislative ambience. From some points of view the fastest and easiest way to do this is by tax cuts. Unfortunately, if tax cuts are temporary they tend to be viewed as windfalls to be saved rather than spent, so that only part of the tax cuts are effectively recycled. Alternatively, if not announced as temporary, tax cuts tend to create a resistance to later tax increases called for by full-employment conditions and large debt-service requirements. This is especially threatening in the present context of political campaigning on the basis of promises of no new taxes. Perhaps the best tax cut would be a cut in the payroll taxes, as promising the maximum proportion of recycling, if this

can be done in the face of outcries that this would be jeopardizing the financial soundness of the social-security system.

Outlays on actual programs, on the other hand, are somewhat harder to start and stop rapidly. There is also the need not to get too far ahead of the effective operation of whatever anti-inflation program is put in place, whether the program of gross markup warrants proposed above or some other, lest anticipatory speculation and inflation get out of hand. The exact program for the start-up period will require careful study.

What happens after the first few months will depend to a large extent on what Keynes called the "animal spirits" of the financial community. At one extreme there could be such horror and alarm at the violation of the conventional wisdom concerning the sinfulness of deficits as to produce a widespread hibernation and flight to foreign shores. More likely, once the financial community has become convinced of the seriousness of the administration's purpose to bring about full employment, and once it is anticipated that demand will shortly use up the spare capacity of existing productive facilities, private capital formation may pick up to the point of absorbing and recycling individual savings sufficiently so that government recycling may for the time being become unnecessary. At the same time, government revenues from increased GNP will increase and outlays for unemployment insurance and welfare will decrease. Also, there may be a need to shut down those governmental programs that compete for real resources with private capital formation, in order to avoid a real "crowding out" (as contrasted with the financial crowding out alleged to occur as a result of government borrowing associated with a tax cut). As a result, a brief period of budget balance or even of surplus may become appropriate.

As the economy hits the ceiling of full employment, however, still another transition becomes necessary. For a while capital formation may continue on its momentum, recycling savings but producing excess capacity that either cannot find labor with which to operate or cannot find markets in which to sell its product. Within a short time after hitting the full-employment ceiling, capital formation will have to drop from that appropriate to a 10-percent growth rate to that suited to a far slower growth rate. At this point attempted savings may again exceed what can be absorbed by private capital formation, even at very low rates of interest. Other ways to recycle the excess will again become necessary, one of which will be renewed government recycling.

XIII. Long-Term Excess of Demand Saving over Private Investment

There is, indeed, no principle of economics that says that there will always be a feasible rate of interest that will equate desired savings and private capital-formation under conditions of steady full employment. Current trends seem to be such as to make such a possibility unlikely. One factor has been a spate of capital-saving innovations and practices. Fiber optics, when fully utilized, costs less per unit of service than previous technologies by orders of magnitude, leaving ductways planned for copper conductors forever surplus; electronic exchanges occupy a fraction of the space formerly required by equivalent electromechanical exchanges; just-in-time practices reduce investment in inventory; improved communications enable more freight to be carried on a single track line with sidings than was formerly carried by a full two-track line; a man assembling electronic gear with a soldering iron uses far less capital than the man in the pulpit of a rolling mill, and service industries generally use less capital per employee than manufacturing, mining, or transportation.

Moreover, before gross investment can begin to recycle private savings, it must first recycle funds set aside in depreciation, amortization, depletion, and obsolescence charges, while rapid obsolescence due to accelerating technological progress makes capital formation relatively insensitive to changes in interest rates. Very low or negative interest rates may stimulate investment in nondepreciating assets such as land, but

even this is limited by the possibility that speculative bubbles may burst, and in any case relatively little recycling is produced thereby, except to the extent that the enhanced asset values cause owners to feel wealthier and spend more.

On the savings side, increased longevity and the high cost of old-age illness lead to increased savings through funded pensions and other provisions for retirement. For this purpose, the lower the rate of interest, the greater is the amount of current savings that must be put aside to provide a given level of retirement security. More recently the increased concentration of income among the very wealthy, who have a high propensity to save, not so much for eventual consumption but largely to accumulate chips with which to play financial games and exercise economic power, has further added to the savings-recycling problem. Some recycling may take place through investment abroad, reflected in a positive trade balance and the production of goods for export, though it is uncertain how far this can be carried in the face of political instability, the danger of creating repayment problems, and the resistance of foreign governments that do not have an effective full-employment policy of their own to our exporting our unemployment to them in this way.

On balance, it may prove impossible, for the foreseeable future, to maintain a steady state of genuinely full employment without a substantial amount of government recycling of savings, a chronic budget deficit, and a long-term increasing trend in the national debt, however distasteful this may be to those ideologically addicted to a balanced budget. It may even prove necessary for the debt to grow at a rate faster than the growth of GNP. The burden of servicing this debt might be kept within bounds by reducing real interest rates, close to zero if need be, though this might imply a higher level of private investment than would be chosen on its own merits. Even contemplating such prospects calls for a significant expansion in our range of habitual thought.

XIV. The Task Before Us

This, then, is the challenge I lay before the economics profession. There is no reason inherent in the real resources available to us why we cannot move rapidly within the next two or three years to a state of genuinely full employment and then continue indefinitely at that level. We would then enjoy a major reduction in the ills of poverty, homelessness, sickness, and crime that this would entail. We might also see less resistance to reductions in military expenditure, to liberalization of trade and immigration policy, and to conservation and environmental protection programs.

I lay before you a plan I believe can accomplish this. It involves government recycling of excess savings plus a method of keeping inflation under control. I believe it can do the job while preserving the essentials of a free-market system. There may be some details to be worked out, but I am confident that the basic concept is sound and workable.

We simply cannot carry on as we have been doing without falling apart as a community and losing what is left of our status of world leadership. If you don't think that something like this can be made to work, then it is up to us to get together to find something that will. Otherwise, if we continue to tie our hands with financial shibboleths and models that tacitly assume a fixed total of resource utilization, we are no better than the feckless castaway whose contribution to the solution of the problem of dealing with cases of canned goods was "let's just assume we have a can-opener."

Trigger Points and Budget Cuts: Explaining the Effects of Fiscal Austerity

By GIUSEPPE BERTOLA AND ALLAN DRAZEN*

We propose and solve an optimizing model which explains counterintuitive effects of fiscal policy in terms of expectations. If government spending follows an upward-trending stochastic process which the public believes may fall sharply when it reaches specific "trigger" points, then optimizing consumption behavior and simple budget-constraint arithmetic imply a nonlinear relationship between private consumption and government spending. This theoretical relation is consistent with the experience of several countries. (*JEL* E62, E21)

Textbook aggregate-demand models suggest simple relationships between the government budget and economic activity: a cut in the government deficit, for example, depressing consumption and output. Though such models heavily influence the design of stabilization policies, their micro foundations are unclear, and their sharp predictions are not always consistent with reality. In many countries, large cuts in government spending carried out as part of stabilization programs have led to *expansions* rather than contractions in economic activity.

The absence of simple relationships between economic variables is often explained theoretically by the importance of expectations of future economic variables, including variables reflecting government policy. The idea that expectations of future policy changes are crucial in understanding seemingly counterintuitive macroeconomic dynamics has been explored before. It formed the basis of the now classic "Some Unpleasant Monetarist Arithmetic" (Thomas J. Sargent and Neil Wallace, 1981). More recently, Drazen and Elhanan Helpman considered the effects of future anticipated fiscal changes meant to balance the budget on current inflation in a closed economy (1990) and the current account in an open economy (1987). Their results indicated that, depending on what sort of policy change was anticipated, any correlation between inflation (or the current account in an open economy) and changes in the budget deficit could be observed.

In this paper we argue that expectations about the discrete character of future fiscal adjustments can help explain the effects of current fiscal policy. The paper develops a model in which significant government spending cuts take place only when the ratio of government spending to output hits a trigger point, so that sharp policy changes occur only infrequently. We use techniques developed in the exchange-rate literature to study such discrete policy interventions. The existence of triggers in a neoclassical model implies a nonlinear relation between the consumption-to-output ratio and the government-spending-to-output ratio. Both the assumptions about the character of fiscal interventions and the implications for the ratio of consumption to output find support in the joint behavior of consumption and government spending observed for several countries. We will look in detail at Denmark and Ireland, where a pattern typical of the relation between private and government consumption in several countries

*Bertola: Department of Economics, Princeton University, Princeton NJ 08544, CEPR, and NBER; Drazen: Department of Economics, University of Maryland, College Park, MD 20742, and NBER. Part of the work on this paper was done at the Innocenzo Gasparini Institute for Economic Research in Milan, which we thank for hospitality. For helpful comments on earlier drafts, we thank Henning Bohn, Elhanan Helpman, Paolo Manasse, Lars Svensson, two referees, and participants in several seminars.

emerges quite strongly: when government spending is a small fraction of GDP, private and government consumption (both relative to GDP) exhibit an inverse relationship, which flattens out as the ratio of government spending to output increases.

The key characteristic of the *expectations view* of fiscal policy is that nonstandard effects of fiscal policy are explained by the role of current policy in shaping expectations of future policy changes. A policy innovation that would be contractionary in a static model may be expansionary if it induces sufficiently strong expectations of future policy changes in the opposite direction. If a cut in government spending induces expectations that future spending *and therefore taxes* will be significantly lower, it may induce an expansion in current private spending. Taken one step further, the effect of increases in government spending may be expansionary if private agents' behavior is based on the expectation that high levels of spending are unsustainable and will soon be cut.

The plan of the paper is as follows. Section I reviews the Danish and Irish experiences. In Section II we present a very simple reference model in which output is constant and government spending follows a Brownian-motion process with positive drift subject to possible discrete adjustment: these simple assumptions, which are suggested by the Danish and Irish experiences, make it possible to obtain a closed-form solution and to illustrate the general phenomena we focus on in this paper. In Section III we propose a solution method for models of stochastic policy changes, and we show that the simple model implies a nonlinear relation between the level of consumption chosen by optimizing agents and the level of government consumption. We discuss the model's implications for consumption behavior and the current account in Section IV. In Section V we discuss the match of the theory to the qualitative features of the experiences of Denmark, Ireland, and other countries and consider how the simple model may be extended to better match observed phenomena. The final section contains concluding comments.

I. Two Examples of Fiscal Stabilizations

Numerous countries undertook fiscal austerity programs in the 1980's. Denmark and Ireland enacted stabilization programs marked by especially sharp cuts in government expenditures during this period. We begin with a very quick overview of these episodes, based on the excellent analysis by Francesco Giavazzi and Marco Pagano (1990).

In the early 1980's the Danish primary government deficit was large relative to GDP, and public debt was growing rapidly. In the autumn of 1982 concerns about the sustainability of the fiscal path led to a decision to undertake a sharp fiscal contraction beginning in late 1982. In the four years that followed, the full-employment primary deficit fell by 10 percent of GDP. The stabilization was expansionary: the consumption-to-GDP ratio rose by several percentage points during 1982–1986, as shown in Figure 1, where we plot the time paths of private consumption c, government consumption g, and current account x (all relative to GDP).[1]

In the early 1980's Ireland was in an even worse position. Giavazzi and Pagano (1990) report a large budget deficit and total national debt equal to 87 percent of GDP. The first attempt to stabilize was undertaken in 1982. By 1984, the (full-employment) primary budget deficit was reduced by 7 percent of GDP, largely via increased taxes. The fiscal contraction had the textbook Keynesian effect, with private consumption falling sharply. A second attempt at stabilization was made in 1987, this time relying on large cuts in government consumption. As in the Danish case, this fiscal retrenchment was quite sharp, with the full-employment budget deficit falling by another 7 percent of GDP between 1987 and 1989. A further similarity to Denmark was the expansionary effect of the fiscal change. The data are presented in Figure 2.

[1]The data are from the International Monetary Fund *International Financial Statistics* tape: lines 96f, 91f, 90c minus 98c, divided by line 99b.

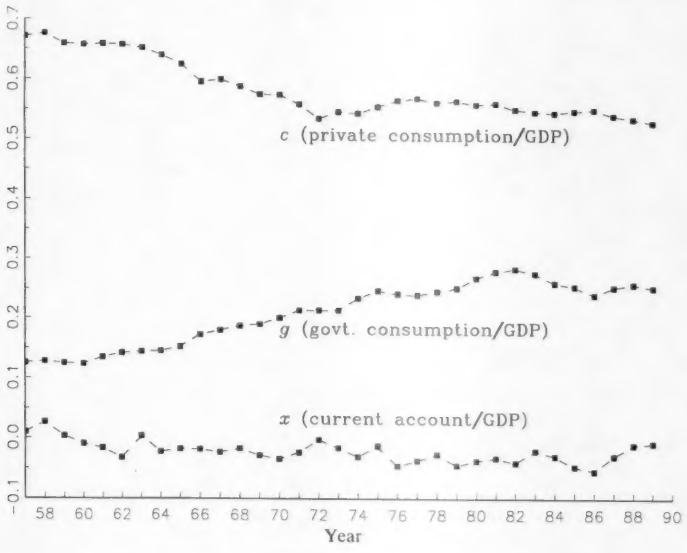


FIGURE 1. DENMARK: TIME SERIES

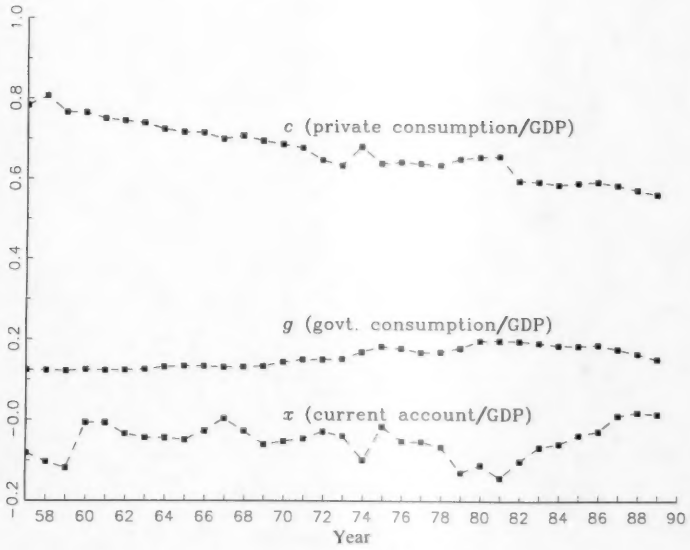


FIGURE 2. IRELAND: TIME SERIES

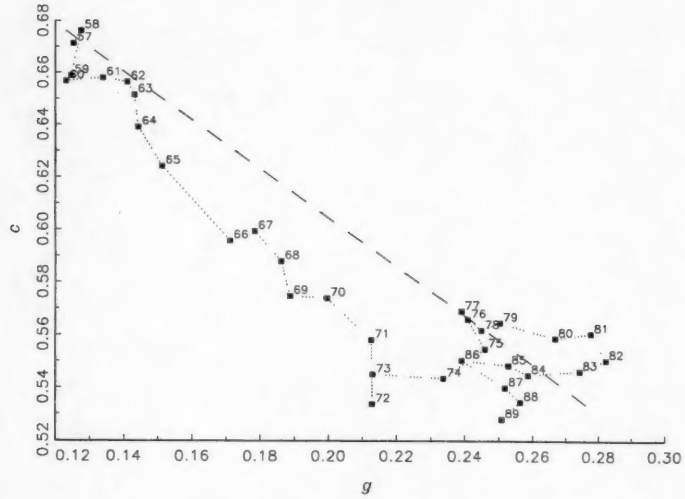


FIGURE 3. DENMARK: PRIVATE AND GOVERNMENT CONSUMPTION RATIOS

Note: Two-digit numbers in the plot denote the years corresponding to each point.

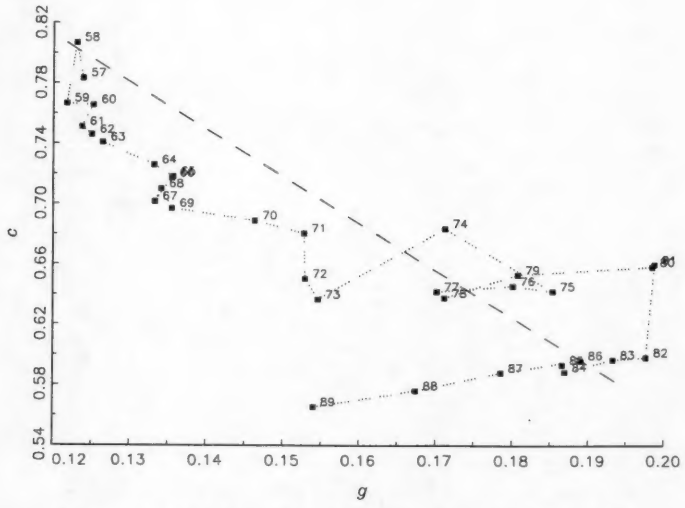


FIGURE 4. IRELAND: PRIVATE AND GOVERNMENT CONSUMPTION RATIOS

Note: Two-digit numbers in the plot denote the years corresponding to each point.

In both countries, we see a similar relation between the private consumption-to-output ratio c and the government consumption-to-output ratio g in the 1957–1989 period, displayed in Figures 3 and 4 for Denmark and Ireland, respectively. (A similar relation emerges if we use government expenditure inclusive of transfers and debt service instead of government consumption.) From 1957 until the early to mid 1970's we see an inverse relation between c and g in both countries; from the early 1970's until the early 1980's c remains flat even though g rises significantly; finally, from 1982–1983 onward there is a significant drop in g with little change in c. Note also that the tax-based stabilization of 1982 in Ireland is associated with a sharp drop in c as g remains constant. The figures suggest, first, that fiscal stabilization takes the form of a sharp retrenchment, following a long upward drift in the ratio of government spending to output; and second, that its effects on private consumption decisions are not as simple as textbook models would suggest.

II. A Simple Model

We consider a nonmonetary small open economy facing a perfect world capital market. Let the world interest rate be r and let this also be the consumer's discount rate. Suppose that all goods are tradable, that output of tradables is fixed, and that the government finances its purchases of tradable goods by lump-sum taxes. (We discuss below the implications of assuming instead that part of government expenditure falls on nontraded goods, or that taxation is distortionary.) Government spending is assumed to follow a known process, which we will specify below. Here, we will relate current consumption of the representative individual to the expected present discounted value of future government spending.

The individual's choice problem may be written as

$$(1) \qquad \max \int_t^\infty e^{-r(z-t)} E_t\{u(C_z)\}\, dz$$

subject to

$$(2) \qquad \int_t^\infty (C_z + \tau_z - Y) e^{-r(z-t)}\, dz \le A_t$$

where $\{C_z\}$ and $\{\tau_z\}$ are, respectively, the consumption and taxation processes; Y is (constant) income; A_t is total private assets, including claims on foreigners and on the government; and $E_t\{\cdot\}$ denotes expectation conditional on the relevant information available at time t.

If government expenditures are stochastic, so are taxes via the government's budget constraint. The first-order condition for the problem in (1) and (2) is

$$(3) \qquad E_t\{u'(C_z)\} = \lambda_t$$

where λ_t is the multiplier on the integral budget constraint. If utility is quadratic in C, the first-order condition implies constant expected consumption over time, $E_t\{C_z\} = C_t$ for all $z \ge t$.[2] Taking expectations on both sides of the integral budget constraint, we obtain

$$(4) \qquad C_t = Y + rA_t - r\int_t^\infty E_t\{\tau_z\} e^{-r(z-t)}\, dz.$$

The government's intertemporal budget constraint implies

$$(5) \qquad \int_t^\infty \tau_z e^{-r(z-t)}\, dz = B_t + \int_t^\infty G_z e^{-r(z-t)}\, dz$$

where $\{G_z\}$ is the government spending process and B_t is government debt (i.e., cumulated past deficits). Substituting (5) into (4) we obtain

$$(6) \qquad C_t = Y + r(A_t - B_t) - r\int_t^\infty E_t\{G_z\} e^{-r(z-t)}\, dz.$$

[2]Our interest is in the inverse relation between current consumption and expected future taxes for a forward-looking consumer. This relation is obviously general; we use quadratic utility to derive a simple functional form.

We note at this point that $A_t - B_t$ equals the country's net foreign assets, since we do not consider capital accumulation. Dividing through by output, we obtain a simple inverse relation between the consumption-to-output ratio and the present discounted value of the ratio of future government expenditure to output, given the ratio f_t of net foreign assets to GDP:

$$c_t = 1 + rf_t - rv_t \tag{6a}$$

where v_t denotes the present discounted value of the ratio of expected future government spending to output; that is,

$$v_t = \int_t^\infty E_t\{g_z\} e^{-r(z-t)} dz. \tag{7}$$

Expectations are unobservable as such. To obtain a relation between (observable) consumption and government spending, we need to specify a process for the latter and to find how expectations of its present discounted value depend on current observables. A number of state variables may be relevant for this purpose. In general, however, any movement in current government spending must imply a change in future spending or taxation to ensure sustainability. To see this, consider a case in which changes in current government spending do *not* give rise to expectations of future offsetting movements: suppose that the ratio of government spending to output, denoted by g_t, follows a random walk (in continuous time, for analytical convenience), so that

$$dg_t = \vartheta\, dt + \sigma\, dW_t \tag{8}$$

where $\{W_t\}$ is a standard Wiener process.[3] We assume the drift ϑ to be strictly positive, reflecting the tendency of government outlays relative to GDP to rise over time in the absence of feasibility concerns. If equation (8) always described the dynamics of $\{g_t\}$, we would have $E_t\{g_z\} = g_t + \vartheta(z - t)$ for all t, all $z \geq t$, and integration of (8) would yield

$$v_t = \frac{g_t}{r} + \frac{\vartheta}{r^2}. \tag{9}$$

Using (9) in (6a), we see that when g displays no tendency to reverse its upward movements, changes in current government spending imply changes in expected future taxation in the same direction and, hence, movements in *current* consumption in the opposite direction.

With g following a random walk, however, the present discounted value of future government spending and taxes would exceed the present discounted value of output with probability approaching 1. Government solvency, however, requires that continuing increases in current government spending must eventually be slowed or reversed. Such stabilizations of an upward trend in the ratio of government spending to output often have a discrete character: there are infrequent sharp changes in government expenditure patterns, with a sharp cut in government spending or its rate of increase occurring only after a significant upward drift, as exemplified by the Danish and Irish experience reviewed above. The infrequency of fiscal regime changes may reflect political constraints blocking agreement on fiscal retrenchment, which can be reached only when the ratio of government expenditure to output reaches levels that are sufficiently high to be deemed critical. (The argument that the enactment of an inevitable stabilization may require the prestabilization situation to deteriorate markedly is explored in Alberto Alesina and Drazen [1991] and Drazen and Vittorio U. Grilli [1993].)

We proceed to model discrete interventions by assuming that the ratio of government spending to output follows the dynamics in (8) but jumps downward discontinuously when it reaches excessively high levels, terming such discrete shifts *stabilizations*. For simplicity, we let all stabi-

[3] Modeling g_t as a (regulated) Brownian motion, we can use the technical apparatus developed in the literature on exchange-rate bands (see the papers in Paul Krugman and Marcus Miller [1992] and Bertola [1993] for a survey). The application of this approach to issues of fiscal sustainability was considered by Miller et al. (1990) and Bertola (1990).

lizations result in g_t dropping to a given common-knowledge level g_s, and we assume that stabilizations may occur either when $g_t = g_c$ or when $g_t = \bar{g}$.[4] Here, $\bar{g}$ represents a ratio of government spending to output sufficiently high to trigger a stabilization with certainty, and g_c is a lower value which the government may or may not consider crucial. We denote the probability of a stabilization when $g_t = g_c$ with p, and we let $p = p_c$ initially. If the stabilization does not take place when g_t reaches g_c for the first time, however, the public recognizes that g_c is not considered "too high" by the government. Thereafter, g_c is no longer deemed critical, and $p = 0$: if the noise component in (8) brings the spending-to-output ratio back to g_c, a stabilization is assigned zero probability. The inevitability of an eventual stabilization follows from the fact that the Brownian-motion dynamics would bring the government's tax obligations to infinity with probability 1. Thus, if g_t goes past g_c, i. eventually drops back to g_s with probability 1 when it reaches $\bar{g}$. At that point, p is reset to its initial value p_c, implying that g_c is once again a candidate for a stabilization. We further assume that there is no analogous feasibility problem for low values of g_t, which are not likely to be observed anyway given the positive drift.

The modeling of stabilizations is not meant to be fully realistic, of course, but we believe that its assumptions do capture politico-economic phenomena of some interest. The public's uncertainty about whether a specific value g_c is high enough to induce an agreement reflects the same constraints that rationalize the existence of trigger points, namely, the difficulty of reaching political agreement on fiscal retrenchment. When no stabilization takes place at a potential trigger point, it is realized that the threat of fiscal unsustainability is not strong enough to induce agreement under current circumstances; if stabilization does occur at g_c, on the other hand, there is no guarantee that the same point will necessarily induce agreement the next time it is hit. The assumption that lower values of g are once again possible critical points after a drastic stabilization at $\bar{g}$ is appealing on theoretical grounds, as it ensures that the economy repeatedly returns to the two possible situations and that (if the economy under consideration were observed for a long enough time) initial conditions would eventually become irrelevant. It may be rationalized considering that, if government spending has been allowed to drift off, this experience may provide more political will not to let it happen again.

[4] More generally, there may be a number of possible trigger points or return points. This simple case illustrates the main results. An alternative approach would be to focus on the inference problem of discovering the true process generating government spending for an observed time series, as suggested in Drazen (1990).

III. Solution

The model outlined above admits a closed-form solution. Stabilizations occur (with given probability) only when the ratio of government expenditures to output reaches "trigger" values, and the dynamics between stabilizations are given by the process (8) which is Markov in levels. The likelihood of stabilizations at every time in the future is then determined, as of time t, by the current level of g_t only, and $\{g_t, p_t\}$ form a bivariate Markov process. All conditional expectations in the integral (7) defining v_t are then functions of g_t and p_t, to imply that $v_t = v(g_t; p_t)$.

By Ito's lemma, whenever $\{g_t\}$ follows Brownian-motion dynamics (and, by our assumptions, p_t is not changing), we have

$$\frac{1}{dt}E_t\{dv\} = \vartheta v'(g_t; p_t) + \frac{1}{2}\sigma^2 v''(g_t; p_t).$$

where the primes denote the partial derivatives of v with respect to g. From the differential form of (7),

$$(7a) \qquad rv_t = g_t + \frac{1}{dt}E_t\{dv_t\}$$

we obtain a differential equation for $v(g_t; p_t)$:

$$(10) \qquad rv(g) = g + \vartheta v'(g_t; p_t) + \frac{1}{2}\sigma^2 v''(g_t; p_t)$$

which must be satisfied by the $v(\cdot)$ function for all values of g for which the probability of a stabilization is zero. The expression in (9) is a particular integral of (10), corresponding to the (unsustainable) case in which stabilizations never occur. All solutions of (10) are then linear combinations of (9) and of solutions of its homogeneous part, in the form

$$h(g) = K_1 e^{\alpha_1 g} + K_2 e^{\alpha_2 g} \tag{11}$$

where α_1 and α_2 solve the characteristic equation

$$\frac{1}{2}\sigma^2\alpha^2 + \vartheta\alpha - r = 0. \tag{12}$$

The roots of this equation are real, distinct, and of opposite sign if $r\sigma^2 > 0$.

The two terms in (11) reflect the upward or downward bias in the discounted expectation of future spending due to the possibility of upward or downward jumps. By the assumptions above, no discrete upward jumps in government spending are triggered by "too low" values of g. Hence, the likelihood of future stabilizations becomes smaller and smaller as g approaches $-\infty$, and it must be the case that

$$\lim_{g \to -\infty} h(g) = 0.$$

This implies that if α_1 is the negative root of (12), then the solution sets K_1 in (11) to zero, while the current probability of stabilization at point g_c will be reflected in the value of the constant of integration associated to the positive α root, which we will denote $K(p)$. Combining (9) and (11), solutions of (10) must then take the form

$$v(g_t; p) = \frac{g_t}{r} + \frac{\vartheta}{r^2} + K(p)e^{\alpha g_t} \tag{13}$$

where

$$\alpha = \frac{-\vartheta + \sqrt{\vartheta^2 + 2r\sigma^2}}{\sigma^2} > 0.$$

The model under consideration allows for two possible values of p_t: p_c, when g_c is a possible stabilization point, and 0, when it is not. At times when a stabilization may occur, $v(g)$ must satisfy the boundary conditions implied by the requirement that the expectation of future government spending not be *expected* to change (i.e., the law of iterated expectations). Consider first the case in which g_t has passed through g_c with no stabilization, so that we are considering the function $v(g_t; 0)$. Since g drops to g_s with probability 1 whenever it reaches $\bar{g}$, the present discounted value of expected future taxes must be identical at $\bar{g}$ and g_s; that is,

$$v(\bar{g}; 0) = v(g_s; p_c). \tag{14a}$$

Similarly, when stabilization at g_c is perceived as possible, there can be no expected jump in $v(g; p)$. Thus, it must be true that

$$v(g_c; p_c) = p_c v(g_s; p_c) + (1 - p_c)v(g_c; 0) \tag{14b}$$

when the expenditure-to-output ratio reaches the critical value g_c and the public is uncertain *ex ante* as to whether a stabilization may occur at that point.

Using (13) and (14), we find that $K(p_c)$ and $K(0)$ must solve the system

$$\begin{bmatrix} e^{\alpha g_s} & -e^{\alpha \bar{g}} \\ e^{\alpha g_c} - p_c e^{\alpha g_s} & -(1-p_c)e^{\alpha g_c} \end{bmatrix} \begin{bmatrix} K(p_c) \\ K(0) \end{bmatrix} = \begin{bmatrix} (\bar{g} - g_s)/r \\ p_c(g_s - g_c)/r \end{bmatrix}. \tag{15}$$

This yields values for $K(p_c)$ and $K(0)$ and, hence, solutions for $v(g_t; p_c)$ and $v(g_t; 0)$ in (13). Quite intuitively, both K values are negative, reflecting the expectation of future downward jumps in the level of g.

To summarize, the changes in future government spending necessary to ensure solvency in response to increases in current government spending imply that a current

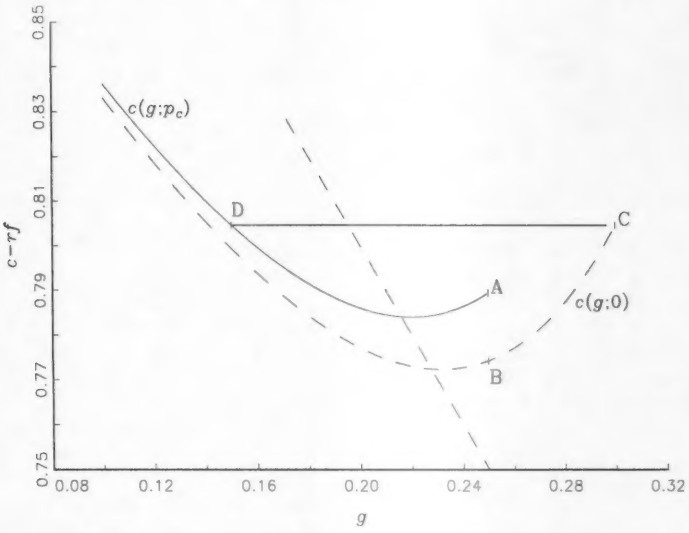


FIGURE 5. MODEL: $r = 0.05$, $\sigma = 0.006$, $\vartheta = 0.0045$, $p_c = 0.5$, $f = 0$

increase in the government spending-to-output ratio, in general, has nonlinear effects on the expected present discounted value of the ratio of future government spending to output. It may therefore be misleading to use simple VAR's to predict future government policies, as was done for example by Giavazzi and Pagano (1990). The mean reversion in government spending which characterizes stabilization has the further implication that the present discounted value of expected future g_t rises less and less as g_t rises. One should note that this nonlinearity does not depend on intervention being discrete. A number of other models would have similar implications: for example, stabilizations may be modeled as discrete changes in the drift ϑ of g_t, or interventions may be infinitesimal rather than discrete. Nor need interventions be infrequent, as long as the present discounted value of future government spending increases less fast than g_t itself. What is crucial is mean reversion, with increases in g making it more likely that it will come back to some return point.

IV. Observable Implications: Consumption and the Current Account

We can now study the relation between observable changes in government spending on the one hand and observable changes in consumption and the current account on the other. Using the results of the previous section we may write (6a) as

$$c_t - rf_t = 1 - rv(g_t; p). \tag{16}$$

Theoretically, f_t is a function of the realized history of $\{g_t\}$; over short intervals, however, f will be roughly constant since changes in f represent the *cumulated* effect of flow consumption decisions from (16). Over longer intervals, changes in f will complicate the relation of c and g in Figure 5 but will not change the underlying characteristic, namely, the flattening out for high values of g, which is the object of our interest. As we do not have information on f, we choose to concentrate on the relationship of c to g, and to work with equation (16) for given f.

We graph the relation given by (16) in Figure 5, choosing parameter values that are realistic for the Danish and Irish experience. The upper, solid curve shows the behavior of consumption for $g_t < g_c$ when there is probability $p_c > 0$ that a discrete policy shift will occur when g_t hits g_c. The lower, dashed curve shows the behavior of consumption when there is no probability of a discrete change in government spending at g_c. The diagonal dashed-dotted line is a 45° line corresponding to a zero trade balance.

Using the solution from (13) for $v(g_t; p)$ in (16) and differentiating, the slope of the relation between c and g is

$$(17) \qquad \frac{\partial c_t}{\partial g_t} = -r\frac{\partial v(g_t; p)}{\partial g_t}$$

$$= -1 - r\alpha K(p)e^{\alpha g}$$

at points where $v(g_t; p)$ is differentiable, that is, at all points other than g_c along $v(g_t; p_c)$ and $\bar{g}$ along $v(g_t; 0)$. [Note that for a change in g occurring in infinitesimal time f is constant to first-order; hence, (17) is strictly correct.]

Several points should be noted. First, if discrete stabilizations were ruled out, movements in g_t would be matched by equal-size movements in c_t in the opposite direction [see equation (9)]: we would then have $\partial c_t / \partial g_t = -1$, and c and g would move along a 45° line in Figure 5. When discrete cuts in spending are included, the relationship of $c - rf$ to g is flatter than a 45° line, and may even slope upward. At low levels of g, c falls less than dollar-for-dollar with increases in current government spending: this of course reflects the fact that increases in current government spending imply increases in the probability of a future discrete spending cut and hence a lower increase in future expected taxes than in the no-expected-intervention case. Thus, higher government expenditure only partially crowds out consumption, and the model—in spite of its extreme neoclassical structure—has distinctly Keynesian observable implications. With fixed output, the difference between the change in consumption and government spending is reflected in the current account.

Consider now the upper curve. When g_c is reached at point A, there is a probability p_c of a discrete cut to level g_s of the ratio of government spending to output. If the cut is realized, we jump to point D: consumption jumps up, reflecting the discrete fall in the present discounted value of future taxes. Note however that consumption rises by less than government spending falls. Alternatively, with probability $1 - p_c$, there is no stabilization when we hit A. This indicates that g_c was not a critical level of spending for the current government, so that expected future spending and thus expected future taxes are discretely higher. Hence consumption falls discretely, as indicated by the jump to point B on the lower curve. That is, a small increase in current government spending may induce a large, discrete fall in consumption. Put another way, the expectation that a stabilization is possible at some perceived critical value will be contractionary if the stabilization does not materialize. Changes in the perceived probability of a discrete cut in government spending will have a similar effect, generating discrete jumps in the consumption-to-output ratio with little change in the ratio of government spending to output.

Along the lower curve, stabilization is expected to occur with certainty when *and only when* the ratio of government consumption to output reaches $\bar{g}$. This curve also bends away from the 45° line, the bend being progressively greater for higher values of g_t. For g_t sufficiently high, increases in g_t induce *increases* in c_t (again, for given f). This is because at high levels of g_t further increases in government spending so increase the probability of a discrete cut in future government spending as to lower the present discounted value of future expected taxes. This result is reminiscent of the inflation-overshooting result found in Drazen and Helpman (1990) and occurs for basically the same reason.

The response of the current account to changes in government spending is, of course, fully determined by the above results since output is fixed. To derive this more formally, we first write the ratio of the current account to output, denoted x_t, as

$$x_t = 1 - c_t - g_t + rf_t. \tag{18}$$

Differentiating with respect to g_t and using (17), we obtain, for values such that $v(g; p)$ is differentiable,

$$\frac{\partial x_t}{\partial g_t} = r\alpha K(p)e^{\alpha g}. \tag{19}$$

Since $K(p)$ is negative this derivative is negative and increasing in g_t, so that along a curve the current account always worsens with increases in g_t. This is identical to the above observation that the $c_t - rf_t$ line bends away from the 45° line, with the bend being greater the larger is g_t.

We quickly note some further features of movements in the current account. At point A, an improvement in the current account occurs with probability 1. If a stabilization occurs (implying a move to D), c_t rises by less than g_t falls, so the current account improves. If no stabilization takes place, we move to B, c_t falls with no change in g_t, and the current account improves. Similarly along the lower curve a stabilization at $\bar{g}$ induces a current-account improvement. Note further that, along the upward-sloping portion of the $c(g_t; 0)$ curve, increases in g_t induce large deteriorations in the current account, as private consumption rises in anticipation of a stabilization.

V. Discussion and Extensions

In this section we compare the implications of the simple model with the experience of Denmark, Ireland, and other countries; to the extent that they do not match, we suggest extensions of the model.

Recalling the evidence reviewed in Section I, our model appears capable of explaining the qualitative features of Danish and Irish consumption data; for realistic parameters, the quantitative match between theory and data is satisfactory as well. Figures 3 and 4 (where dashes plot a 45° line) display a nonlinear relation between the ratio of private consumption to output and the ratio of government consumption to output which is quite similar to the theoretical one in Figure 5. The main puzzling feature of the Danish and Irish data, namely, the sharp drop in government spending with no similar change in private consumption at the time of stabilization (or the "expansionary effects of stabilization"), is readily explained by the model on the assumption that the stabilization was anticipated.

Our theoretical model was kept simple in order to make clear the general insight of the expectations approach to fiscal policy: solvency of the government and forward-looking consumer behavior imply that, in a neoclassical model, current consumption will fall less than one-for-one with increases in current government spending because changes in current government spending engender expectations of future fiscal changes in the opposite direction. As mentioned above, other models of the process generating government spending would have similar implications, as long as they maintained the characteristic of mean reversion that fiscal sustainability implies.

The simple model we present has another feature that is consistent with the data, namely, the effect of changes in the perceived probability of a discrete cut in government spending. The 1982 drop in consumption in Ireland is often attributed to the Keynesian effects of the sharp increase in taxes with the first stabilization attempt (Giavazzi and Pagano, 1990). Our model suggests how expectational effects could explain this drop. It has been argued (Rudiger Dornbusch, 1989) that the need for fiscal austerity was recognized already in 1981. In our framework, then, the crucial event in 1982 was not the tax increase (which in a neoclassical model should have little effect on current consumption) but the *absence* of a significant cut in g when this had come to be expected. According to the

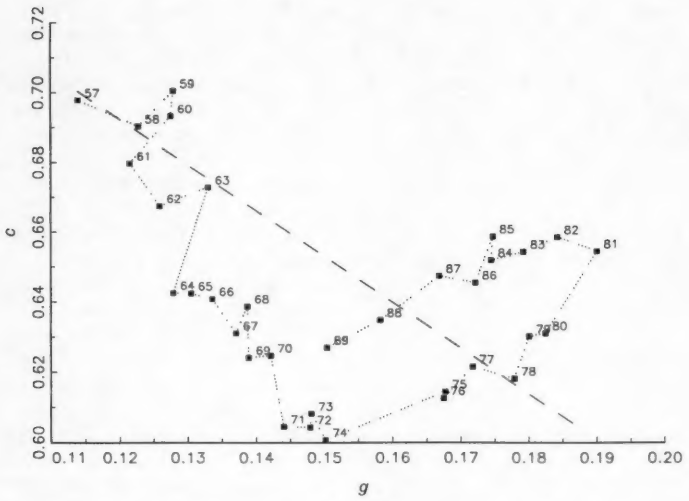


FIGURE 6. BELGIUM: PRIVATE AND GOVERNMENT CONSUMPTION RATIOS

Note: Two-digit numbers in the plot denote the years corresponding to each point.

model, the failure of the cut to materialize at what had been perceived as a critical point should and did result in a drop in c with no change in g, and in a "failed" stabilization.[5]

Similar patterns of an inverse relation between c and g for low values of g which flattens out as g rises are observed elsewhere, for example, in Belgium (Fig. 6), Sweden (Fig. 7), and the United Kingdom (Fig. 8), all of which enacted programs of fiscal austerity in the 1980's. (The dashed line is again a 45° line.)

Plotting the ratio of the current account to GDP against g using Irish and Danish data yielded a significantly worse match of the model to the data than the plots of c against g (plots not shown). We will briefly discuss the possible reasons for the poor fit between the simple model above and the data.

First, the relation derived in (19) arises because of the accounting identity (18). Changes in the sum of c and g must be mirrored by changes in x in the opposite direction when output is constant. If investment were added as the final component of output, the relation in (19) would apply to the derivative of the *sum* of investment and the current account (relative to GDP). The relation of this sum to g observed in the data for our sample of countries (not shown here) is indeed closer to that predicted by the model but contains no new information. Of course, this raises the issue of extending the model to include investment behavior. Changes in the anticipated fiscal environment should strongly influence current investment decisions, as for example the current investment boom in Mexico suggests. The precise effect of future fiscal policy on the forward-looking investment decisions of firms, however, would depend crucially on the nature of the adjustment costs they face. We plan in future research

[5]For Denmark, our model would similarly associate the sharp vertical drop in the consumption-to-GDP ratio in 1972 with a fall in the perceived probability of a government spending cut; we do not know whether this was indeed the case.

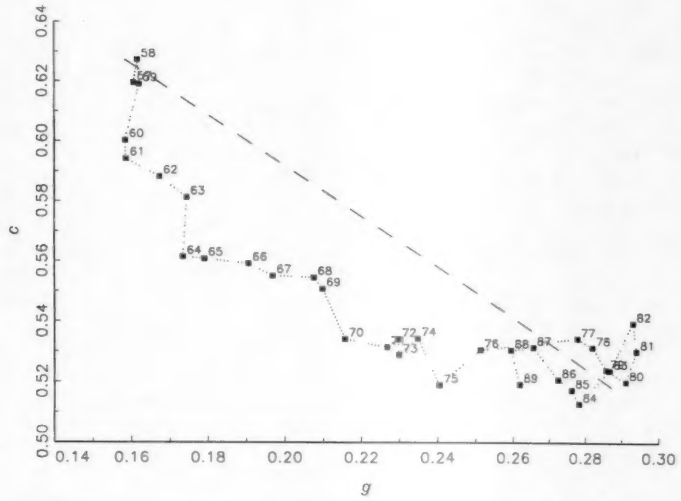


FIGURE 7. SWEDEN: PRIVATE AND GOVERNMENT CONSUMPTION RATIOS

Note: Two-digit numbers in the plot denote the years corresponding to each point.

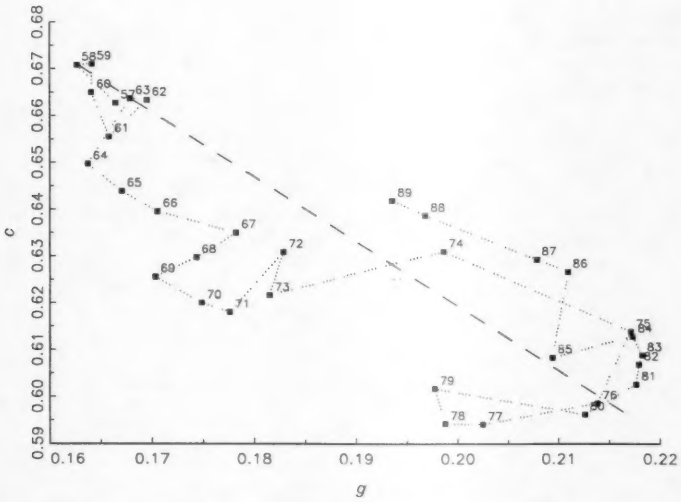


FIGURE 8. UNITED KINGDOM: PRIVATE AND GOVERNMENT CONSUMPTION RATIOS

Note: Two-digit numbers in the plot denote the years corresponding to each point.

to model forward-looking investment behavior along with consumption decisions. The point of this paper, however, was to emphasize the general message of how expectations of future fiscal shifts affect consumption in a nonlinear way, not to set out a complete macro model. The effects of expected fiscal policy on consumption that we focus on would not be qualitatively changed by explicit consideration of investment.

Second, a distinction between traded and nontraded goods could also help explain the poor fit of theory and current-account data. If we added nontraded goods to the model and assumed that the government purchased both types of goods, the results for consumption would be similar, but those for the current account could change significantly. As with tradables, an increase in government expenditure on nontraded goods would increase the probability of a future cut in government consumption of nontraded goods. With fixed output in each sector, market clearing implies that current private consumption of nontraded goods would fall by exactly the amount that government consumption rose. Hence the sign of the consumption results would be unaffected, except perhaps in the upward-sloping region of the curve. For the current account, however, a future anticipated cut in government consumption of nontraded goods could induce a surplus rather than a deficit for any value of g_t if traded and nontraded goods were complements (see Drazen and Helpman, 1987). That is, with fixed output, the spending cut implies an increase in consumption of nontraded goods and thus, via complementarity, increased consumption of tradables *on the date the cut takes place*. This effect taken alone implies that a current-account surplus will be run in anticipation of a stabilization in order to finance higher future expected consumption of tradables. The net effect might therefore be ambiguous.

A further problem related to the poor match between theory and data is that, while the nondistortionary tax model implies that the relation between c and g should be everywhere flatter than the 45° line, the data often display a steeper relation (for example in Denmark in the 1960's and early 1970's). One way to explain this is that net foreign assets are falling over time, so that permanent income is below the present discounted value of the endowment flow minus the present discounted value of future taxes. A second mechanism that might explain this feature of the data is distortionary taxation. We therefore explore briefly how the theoretical results would be affected if taxes were distortionary and current output depended negatively on the current level of distortionary taxation.[6]

Consider first an increase in current government spending with no increase in current taxes. Current output is unaffected, while current consumption will be "doubly" affected, relative to the earlier model. The increase in current government spending raises the expected value of future government spending and hence taxes. This reduces consumption for a given level of permanent income, as before. In addition it reduces permanent income due to the distortionary nature of (future) taxation. Hence, the fall in the ratio of consumption to output in response to an observed increase in the ratio of government spending to output will be greater than in the earlier model (or the rise will be smaller if we are on the upward-sloping portion of the curve). If this distortionary effect were strong enough, the relation would be steeper than the 45° line. A discrete cut in government spending would similarly cause a greater consumption boom in the case where it was not expected with certainty.

At the other extreme, consider an increase in current taxes with no change in current government spending and hence ex-

[6]Alternatively, current output could depend on the expected level of future distortionary taxation, which would arise from factor-accumulation decisions or intertemporal substitution in factor supply, suggesting the importance of carefully modeling investment. We believe this to be an important possibility but leave a formal analysis for a later paper. Paolo Manasse (1991) extends the model proposed here to allow for tax-smoothing behavior on the part of the fiscal authorities.

pected future spending, and suppose that the current level of taxes gives no information about the future level of taxes not already contained in government spending. In the previous model with Ricardian equivalence, this had no effect. In this model, total consumption C_t is still unaffected, since permanent income is unaffected. There is, however, a transitory fall in income and hence a rise in the ratio c_t. Therefore, a fall in the consumption-to-GDP ratio, such as we saw in Ireland in 1982 after the tax-based stabilization attempt, requires a fall in the perceived probability of government spending cuts in both the distortionary tax case and the nondistortionary tax case. An increase in distortionary taxes would imply an unambiguous worsening of the current-account-to-GDP ratio.

To summarize, we would argue that the general implication of letting taxes be distortionary in a neoclassical model is that reductions in the deficit via tax increases can have real effects, effects different than those arising from deficit reductions due to government spending cuts. In the latter case, private consumption will increase, while output will at worst remain unchanged and may rise if current output depends negatively on future expected taxes. In the former case of tax increases with no change in expected government spending, output will fall, and consumption will at best remain unchanged and may fall if a government spending cut had been expected. The expectations view of fiscal policy therefore appears to be capable of explaining the very different effects of government spending cuts and tax increases which Giavazzi and Pagano (1990) found to characterize numerous countries in the 1980's.

VI. Conclusions

We proposed a model in which the effects of fiscal policy on current economic activity depend on what sort of expectations of future policy are engendered, focusing in particular on the consumption effects of expected discrete changes in future fiscal policy. A simple model in which the ratio of government spending to GDP follows regulated Brownian motion predicts a nonlinear relation between consumption and government spending, with jumps in the ratio of consumption to GDP at certain points. We observe this type of nonlinear relation between the ratio of consumption to GDP and the ratio of government spending to GDP in Denmark, Ireland, and several other countries. Other specifications for the process of government spending that capture the qualitative nature of stabilizations would have similar effects.

We believe that the results we obtain are more general than the specific application we present and that the simple model in this paper should be read as an illustration of a general phenomenon. The results indicate several directions for further research.

The expectations view of fiscal policy indicates, among other things, that a neoclassical model may generate observable implications that look quite Keynesian. In a nonmonetary closed economy with fixed output, an increase in government spending that induces the expectation of a future cut will cause interest rates to rise. If effort were endogenous, increases in g would imply expansions in output as consumers attempt to increase current income. It appears feasible to specify and solve a two-country model in which increases in g at home induce increases in world interest rates and world output and a deterioration of the home country's trade balance. On the empirical side, verification of the role of net foreign assets and distortionary taxation is important.

It would of course be beneficial to develop a way of formally estimating and testing a model stressing fiscal expectations. Unfortunately, however, this is very difficult, or even impossible, given the infrequent nature of stabilizations or drastic changes in fiscal regimes and the short data span. However, the model does deliver predictions that are in broad agreement with the experiences of a number of countries and, on a more general methodological level, points out possible problems with applying

simple time-series methods when economic considerations suggest significant nonlinearities in behavior.

REFERENCES

Alesina, Alberto and Drazen, Allan, "Why Are Stabilizations Delayed?" *American Economic Review*, December 1991, *81*, 1170–88.

Bertola, Giuseppe, "Discussion of Miller, Skidelsky, and Weller," in Rudiger Dornbusch and Mario Draghi, eds., *Public Debt Management: Theory and History*, Cambridge: Cambridge University Press, 1990, pp. 311–14.

_____, "Continuous-Time Models of Exchange Rates and Intervention," in Frederick van der Ploeg, ed., *Handbook of International Macroeconomics*, London: Blackwell, 1993 (forthcoming).

Dornbusch, Rudiger, "Credibility, Debt and Unemployment: Ireland's Failed Stabilization," *Economic Policy*, April 1989, *8*, 174–209.

Drazen, Allan, "Discussion of Giavazzi and Pagano," in Olivier Blanchard and Stanley Fischer, eds., *NBER Macroeconomics Annual 1990*, Cambridge, MA: MIT Press, 1990, pp. 117–22.

_____ **and Grilli, Vittorio U.,** "The Benefits of Crises for Economic Reform," *American Economic Review*, 1993 (forthcoming).

_____ **and Helpman, Elhanan,** "Stabilization with Exchange Rate Management," *Quarterly Journal of Economics*, November 1987, *102*, 835–55.

_____ **and** _____, "Inflationary Consequences of Anticipated Macroeconomic Policies," *Review of Economic Studies*, January 1990, *57*, 147–66.

Giavazzi, Francesco and Pagano, Marco, "Can Severe Fiscal Contraction be Expansionary? Tales of Two Small European Countries," in Olivier Blanchard and Stanley Fischer, eds., *NBER Macroeconomics Annual 1990*, Cambridge, MA: MIT Press, 1990, pp. 75–110.

Krugman, Paul and Miller, Marcus, eds., *Exchange Rates and Currency Bands*, Cambridge: Cambridge University Press, 1992.

Manasse, Paolo, "Are Taxes Too Low?" I.G.I.E.R. Working Paper No. 4, Università Bocconi (Milan), 1991.

Miller, Marcus, Skidelsky, Robert and Weller, Paul, "Fear of Deficit Financing: Is It Rational?" in Rudiger Dornbusch and Mario Draghi, eds., *Public Debt Management: Theory and History*, Cambridge: Cambridge University Press, 1990, pp. 293–310.

Sargent, Thomas J. and Wallace, Neil, "Some Unpleasant Monetarist Arithmetic," *Federal Reserve Bank of Minneapolis Quarterly Review*, Fall 1981, *5*, 1–17.

Economic Policy, Economic Performance, and Elections

By JOSEPH E. HARRINGTON, JR.*

This paper explores the role of reelection pressures in determining economic policy. By assuming that voters are uncertain as to the efficacy of different policies, a reelection process is derived which depends on both a politician's past policies and the level of economic activity during his time in office. It is shown that the less uncertain are voters as to which policy is best, the more likely is a politician to manipulate policy for reelection purposes. Manipulation entails implementing the policy that is more likely to be well received rather than the one that maximizes income. (*JEL* D78, E60)

The president is a major player in the U.S. economy. His decisions with respect to fiscal and monetary policy, the regulation of markets, the enforcement of antitrust law, and the erection of trade barriers, for example, can have sizable economic effects. In light of the significance of presidential economic policy, it is important to understand the forces that underlie these policies. One factor that has drawn the attention of both economists and political scientists is the pressures of reelection. In principle, a first-term president can influence his reelection prospects directly through the policies he selects and indirectly through the impact of these policies on economic performance. A president who strategically manipulates policy in order to improve his reelection prospects not only creates a policy bias in the current period but also affects future policy if he is successful in influencing the outcome of the election.

In order to explore the incentives for the manipulation of policy, one must first develop a model of the reelection process. The two leading lines of research follow from the pioneering work of Anthony Downs (1957) and V. O. Key, Jr. (1966).[1] Downs argued that an incumbent's success in bidding for reelection is largely determined by his past policies and actions. That is, voters look at the past decisions of the incumbent and make inferences about the decisions he will make in the future. In contrast, Key argued that voters largely ignore the past decisions of politicians and instead focus on measures of performance; in particular, economic performance.

Recent related research in economics and political science can be placed in either of these two camps. Past policies are modeled as the dominant factor in elections in the work of Kenneth Rogoff and Anne Sibert (1988), Rogoff (1990), W. Robert Reed (1989a, b), and Gregory D. Hess (1991). For example, Rogoff and Sibert (1988) and Rogoff (1990) assume that the incumbent's level of competency is private information, where greater competency means an ability to generate higher revenue or a higher level of government services. The incumbent chooses monetary policy (Rogoff and Sibert,

* Department of Economics, Johns Hopkins University, Baltimore, MD 21218. I thank Dan Klein, Stergios Skaperdas, two anonymous referees, and seminar participants at the University of Virginia for their comments and suggestions. Any remaining errors are, of course, my own. This paper was completed while the author was on leave at Northwestern University. The hospitality of the Department of Management and Strategy is greatly appreciated.

[1] Early research includes William D. Nordhaus (1975) and C. Duncan MacRae (1977) on the political business cycle and Morris P. Fiorina (1981) and Douglas A. Hibbs, Jr. (1987), which build on Key's idea of retrospective voting. Also, see M. Daniel Bernhardt and Daniel E. Ingberman (1985), Alex Cukierman and Allan H. Meltzer (1986), and John R. Lott, Jr. and W. Robert Reed (1989). In contrast to work to be reviewed later, these analyses do not formally model voters so that voting behavior must be specified ad hoc.

1988) or fiscal policy (Rogoff, 1990) and thereby influences some measure of economic performance. The outcome of the election is shown to depend only on the incumbent's past policies and is independent of economic performance.

A second line of research follows the lead of Key (1966). John A. Ferejohn (1986), David Austen-Smith and Jeffrey S. Banks (1989), and Alberto Alesina and Alex Cukierman (1990) specify that the incumbent's actions or policies are not observed by the electorate. As a case in point, Alesina and Cukierman (1990) assume that the incumbent's policy preferences are private information. Since observed economic performance is a noisy signal of the incumbent's underlying policy, voters use performance to draw inferences about the incumbent's current policy and thus about his future policy intentions. However, even in this class of theories, if the incumbent's policies were observable then voting would depend only on policy and thus be independent of performance.[2]

Though there is minimal empirical work that takes account of the electoral impact of a president's policies, there is considerable empirical support for the hypothesis that the performance of the economy affects the outcome of the election. Summarizing the evidence, D. Roderick Kiewiet and Douglas Rivers (1985) find that a 1-percent decline in real income is associated with a reduction of the incumbent party's vote share of between 0.5 percent and 1 percent. In interpreting this empirical regularity, existing theories would argue that the performance of the economy is simply acting as a proxy for the president's economic policies. The implication is that if the relevant policy instruments were observable and controlled for in the empirical specification, the significant role of economic performance would disappear.

In this paper, I develop a theory that provides an alternative explanation for the empirical relationship between economic performance and electoral outcomes. It is argued that measures of economic performance are of intrinsic interest to voters and are not simply acting as proxies for the president's economic policies. An electoral process is endogenously derived which depends on both the (observable) economic policies of the incumbent *and* the performance of the economy during his time in office. The key modeling innovation is to allow voters to be uncertain as to which economic policy is more productive in generating income. Furthermore, voters and politicians can have different opinions as to which policy is best. Thus, voters are uncertain about which policy is best *and* about the incumbent's future policy intentions, as they are uncertain of which policy the incumbent believes is best. We might add that preceding research has ignored this source of voter uncertainty, as voters' policy preferences are treated as primitive. In fact, policy preferences are derived from one's preferences on income *and* one's beliefs over the effectiveness of different policies. An important implication is that voters' policy preferences can change over time in response to new information about the efficacy of different policies. The endogeneity of voters' policy preferences will play a central role in the analysis.[3]

One of the virtues of such a theory is that it can provide insight as to when elections will be largely influenced by policy and when they will be largely influenced by performance. We find that voting behavior and thereby electoral outcomes are more sensitive to the policy chosen, the more confident are voters in their beliefs as to which policy is best. Voting behavior approaches being performance-based, as in Alesina and

[2]Banks and Rangarajan K. Sundaram (1990) consider a model of adverse selection in which incumbents differ according to their competency but do not make decisions. Obviously, performance plays an important role in that setting.

[3]Though, to the best of my knowledge, this is the first electoral model to allow for uncertainty over the mapping from policies to outcomes, this type of uncertainty has been allowed for in legislative models (see Austen-Smith and William Riker, 1987; Thomas W. Gilligan and Keith Krehbiel, 1989).

Cukierman (1990), when voters become more uncertain as to which policy is best, and it approaches being policy-based, as in Rogoff (1990), when voters become less uncertain.

Whether or not electoral outcomes are dominated by policy or performance turns out to be an important determinant of the policy decisions of elected officials. My analysis shows that the more sensitive are electoral outcomes to policy, the greater is the extent of policy manipulation by the incumbent for the purposes of reelection. Specifically, the incumbent is more likely to implement the policy that will be well received and less likely to implement the policy that will maximize income. This behavior not only creates a bias in current policy but also results in less than desirable incumbents being reelected, thereby influencing future policy.

I. The Model

Consider an elected office in which it is the responsibility of the officeholder to select and implement a policy that affects the income of his constituents. I assume that the elected official chooses from among just two policies, denoted policy 1 and policy 2. While this assumption is made for reasons of tractability, it has some appeal in that voters may find it difficult to make subtle distinctions between policies. For example, they may only take note as to whether or not taxes are raised. In this sense, policies may be quite broadly defined. In fact, they may be so broadly defined as to be more appropriately interpreted as ideologies; for example, extensive or minimal government intervention in the economy.

Implementation of policy x, where $x \in \{1,2\}$, generates per capita income $y = \omega_x + \varepsilon$ where ω_x is the policy-specific component and ε is a random shock. All voters receive income y, so differential policy effects are not issues. It is assumed that ε is independently and identically distributed across time and policies and that $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ where $\sigma_\varepsilon^2 > 0$ and finite. After voters observe the policy and the realized level of income, the incumbent runs for reelection against a challenger. The winner implements a policy, and the game ends after income is realized. With the exception of assuming that voters observe policy, this structure is identical to that specified in Alesina and Cukierman (1990).

An incumbent is presumed to care about holding office and about the level of income, the latter because he is altruistic or his income is also y. The incumbent is risk-neutral with respect to income, and his utility is additively separable in income and holding office. Let k denote the value attached to holding office, referred to as "ego-rents" by Rogoff and Sibert (1988). The utility to the incumbent from being in office when income y is realized is then represented by $\theta y + k$ where $\theta \geq 0$ and $k > 0$. His total payoff for the two-period game is simply the sum of the discounted single-period utilities where $\delta \in [0,1]$ is the discount factor. Voters are assumed to care only about income and are risk-neutral as well. The horizon is such that each voter will want to maximize his expected income so that voters' time preferences do not play a role.

A singular feature of the model is that I do not take voters' policy preferences as being primitive. Instead I specify voters' preferences over outcomes and their beliefs over the efficacy of policies 1 and 2. Voters are allowed to be uncertain as to which policy is more effective; this uncertainty is represented by having the policy-specific parameters ω_1 and ω_2 be unknown. Though the income generated by a policy is the same across voters, voter heterogeneity in their policy preferences arises by allowing voters to have different beliefs as to the efficacy of policies 1 and 2. Voters' beliefs on ω_1 and ω_2 are assumed to be normally distributed where the prior expectations on ω_1 and ω_2 are allowed to vary within the electorate. Thus, a voter of type (μ_1, μ_2) is specified to have beliefs: $\omega_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ and $\omega_2 \sim \mathcal{N}(\mu_2, \sigma^2)$ where $\sigma^2 > 0$ and finite. Each voter's beliefs are private information.

The significance of endogenizing policy preferences is that policy preferences are not constrained to be fixed over time. In

particular, if voters receive new information about the efficacy of either or both policies, their beliefs on (ω_1, ω_2) can change and, along with them, their policy preferences. The endogenous nature of voters' policy preferences will play a central role in the analysis.

In specifying the type-space of voters, one needs to consider the ways in which voters' beliefs might differ. Two sources of heterogeneity in voters' beliefs become immediately apparent. First, voters could have different beliefs as to which policy will generate higher income on average. Second, voters could differ in terms of their optimism; that is, one voter could expect income to be higher under both policies than does another voter. In order to prevent too many sources of voter heterogeneity from confounding the analysis, I have decided only to allow voters to differ in their beliefs over which policy is best. I chose it because I believe it is the more important of the two.

Following through on the preceding discussion, the type space must be specified so that no voter type is larger than another voter type in both components. This means that the type space is represented by a decreasing function in $\mathbb{R}^2$ so that it is of the form $\{(\mu, \phi(\mu)) | \mu \in \mathrm{A}\}$ where $\phi(\cdot)$: $A \rightarrow B, \phi(\cdot)$ is a strictly decreasing function, and $\mathrm{A}, \mathrm{B} \subset \mathbb{R}$. I will specifically assume that $A = B = [0,1]$ and $\phi(\cdot)$ is linear. Thus, the type space is of the form $\{(\mu, 1-\mu) | \mu \in [0,1]\}$ so that a single parameter represents a voter's type. I conjecture that all results are robust to allowing $\phi(\cdot)$ to be any differentiable decreasing function.

To summarize, since $y = \omega_x + \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, all voters believe that income is normally distributed with variance σ_ε^2. Where voters disagree is over expected income (denoted ω_x for policy x). A type-μ voter's beliefs over expected income from policy 1 are represented by $\mathcal{N}(\mu, \sigma^2)$ and from policy 2 they are represented by $\mathcal{N}(1-\mu, \sigma^2)$. Note that a type-$\mu$ voter initially believes policy 1 to be more effective than policy 2 if and only if $\mu > \frac{1}{2}$.

Like voters, politicians have beliefs over (ω_1, ω_2) which are private information. In contrast to voters, however, the politician is assumed to be rather dogmatic in that a type-α incumbent believes with certainty that $(\omega_1, \omega_2) = (\alpha, 1-\alpha)$, $\alpha \in [0,1]$. Hence, a type-α incumbent believes that income is normally distributed with mean α when policy 1 is implemented and mean $1-\alpha$ when policy 2 is implemented, with a variance of σ_ε^2. This specification of politicians' beliefs considerably simplifies the analysis and allows me to focus on the implications of voters' policy preferences being endogenous. What is critical to the analysis is that the policy that a politician believes to be most effective does not change over the course of his time in office. At a minimum, it seems plausible that voters are relatively more uncertain than politicians about the effectiveness of different policies.

Since each agent's type is private information, it is necessary to specify each agent's beliefs over the other agents' types. To begin, think of the electorate as being defined over the type space $[0,1]$. Assume that there is either a finite and odd number of voters or a continuum of voters so that the median voter exists. Later analysis will show that the median-voter theorem applies so that the incumbent need only be concerned with the beliefs of the median voter. It is assumed that the median voter's type is randomly drawn from $[0,1]$ according to the differentiable cumulative distribution function $G(\cdot)$: $[0,1] \rightarrow [0,1]$. Thus, the incumbent's prior beliefs on the median voter's type are represented by $G(\cdot)$. The incumbent's type as well as the challenger's type are independent random draws from $[0,1]$ chosen according to the continuous cumulative distribution function $F(\cdot)$: $[0,1] \rightarrow [0,1]$. Note that a type-α candidate believes that policy 1 is more effective if and only if $\alpha \in (\frac{1}{2}, 1]$. Let $f \equiv 1 - F(\frac{1}{2})$ so f is the prior probability that a candidate believes that policy 1 is more effective. I assume that $f \in (0,1)$. Voters' prior beliefs on a politician's type and the incumbent's prior beliefs on the challenger's type are represented by $F(\cdot)$.

The key features of the informational structure are that voters and politicians have different beliefs over the effectiveness of

different policies *and* that this is common knowledge. Under the common prior assumption, these assumptions are inconsistent (Robert J. Aumann, 1976). Hence, a rationale for the informational structure is that agents are endowed with distinct prior beliefs. The heterogeneity in prior beliefs may be due to genetic or cultural differences (e.g., the biases of one's parents) but are not the result of differences in information. While agents differ in terms of the formation of their priors, they all receive the same information over time so that no agent believes he can learn anything from what another agent believes. While this is clearly an abstraction, what is important is that it is possible for agents to know that their beliefs are different and yet have no inclination to adjust them. Perhaps it is more appropriate to refer to these beliefs as "opinions," since people regularly agree to having different opinions.[4] Semantics aside, this structure is admittedly unorthodox. Nevertheless, I have chosen to explore it because I believe that its features are fundamental and prevalent in many economic and political settings. At a minimum, it is a quite plausible structure and would appear to be consistent with some observed behavior (e.g., betting on sporting events between individuals).[5]

In concluding this section, I will describe the form of players' strategies. To begin, the strategy of the incumbent tells him which policy to implement in his first term given his beliefs. Hence, the incumbent's strategy is of the form $x(\cdot)$: $[0,1] \rightarrow \{1,2\}$. I do not formally model the policy choice of the elected official in the second period, as it is known that optimality requires that he implement the policy which he believes will maximize expected income.[6] The decision faced by a voter is whether to cast his vote for the incumbent or for the challenger. At the time he casts his vote, a voter knows his own beliefs on (ω_1, ω_2), the policy chosen by the incumbent, and the realized level of income. A voter's strategy is then of the form $t(\cdot)$: $[0,1] \times \{1,2\} \times \mathbb{R} \rightarrow \{I, C\}$ where I denotes "incumbent" and C denotes "challenger." To break indifference, I assume that the voter who is indifferent between the incumbent and the challenger will vote for the incumbent. This is easily justified by assuming that there are some small benefits from seniority or some small transactions costs associated with transition.

II. Optimal Voting Behavior

I begin by characterizing a perfect Bayesian equilibrium in which the incumbent implements a policy if and only if he finds it sufficiently effective. In this section, I derive optimal voting behavior when the incumbent uses such a policy rule. In Section III, I show the existence of such an equilibrium and characterize the bias in policy created by reelection pressures. Then, in Section IV, I derive necessary and sufficient conditions for this equilibrium to be the unique solution and go on to explore the determinants of policy.

When the incumbent implements a policy if and only if he finds it to be sufficiently effective, his strategy takes the following general form, where α^* is an undetermined variable:

$$(1) \qquad x^*(\alpha) = \begin{cases} 2 & \text{if } \alpha \in [0, \alpha^*] \\ 1 & \text{if } \alpha \in (\alpha^*, 1]. \end{cases}$$

For now suppose that $\alpha^* \in (0,1)$. Recall that a type-α incumbent expects policy 1 (2) to generate income of α $(1-\alpha)$. Hence, the

[4]The prevalence of differences of opinions among economists over optimal economic policies is documented in Bruno S. Frey et al. (1984).

[5]In models of rational agents, it is standard to assume that an agent knows all of the implications of his knowledge (see Jakkoo Hintikka, 1975; Barton L. Lipman, 1991). When this is not true, it is possible for an agent to believe something that is logically inconsistent with his body of knowledge. In that case, it is easy to imagine it being common knowledge among agents that they hold different beliefs, for they could all derive different implications from the same information.

[6]The optimal policy for the elected official in the second period is well defined, generically, when $\theta > 0$. When $\theta = 0$, I assume that candidates have lexicographic preferences over tenure and income.

higher is α, the more effective the incumbent finds policy 1 vis-à-vis policy 2.

Voters conjecture that the incumbent uses the policy function in (1). Thus, in response to observing that the incumbent implemented policy 1 in his first-term, voters will infer that the incumbent's type lies in $(\alpha^*, 1]$. Recall that, if reelected, the incumbent will implement policy 1 if and only if $\alpha > \frac{1}{2}$ (i.e., he will act to maximize expected income in his final term in office). As a result, voters assign probability $[1 - F(\max\{\frac{1}{2}, \alpha^*\})]/[1 - F(\alpha^*)]$ to the incumbent implementing policy 1 if reelected, given that he implemented policy 1 in his first term. Note that no information is revealed about second-term policy when $\alpha^* = 0$ (as all incumbent types implement policy 1) so that voters' posterior probability is the same as their prior. Alternatively, if $\alpha^* > \frac{1}{2}$ then voters know for certain that, if the incumbent implemented policy 1 in his first term, he will do so again if reelected.

Since voters assign probability $1 - F(\frac{1}{2})$ to the challenger implementing policy 1, implementation of policy 1 in his first term indicates that the incumbent is more likely than the challenger to implement policy 1 in the future. As a result, when $x = 1$, a voter votes for the incumbent if and only if he desires policy 1 to be implemented again. By a similar argument, one can show that when $x = 2$, a voter votes for the incumbent if and only if he prefers policy 2. Therefore, when the incumbent uses the policy rule $x^*(\cdot)$, voters tend to associate the incumbent's future policy intentions with his past policies. This is reflected in voting behavior in that a voter votes for the incumbent if and only if he desires to have the same policy implemented again.

Having just derived voters' posterior beliefs on the incumbent's future policy intentions, the next step is to derive their posterior beliefs on the effectiveness of different policies. Suppose that policy x is implemented and y is realized. Since $y = \omega_x + \varepsilon$, income is then a noisy signal of ω_x. Let $E[\omega_x | y, \mu]$ denote a type-μ voter's posterior expectation of ω_x when policy x was implemented and income was y. Under the normality assumptions, Bayes' Rule yields the following simple form (Morris H. DeGroot, 1970):

$$(2)\quad E[\omega_1 | y, \mu] = \rho y + (1-\rho)\mu$$

$$(3)\quad E[\omega_2 | y, \mu] = \rho y + (1-\rho)(1-\mu)$$

where $\rho \equiv \sigma^2/(\sigma^2 + \sigma_\varepsilon^2)$.

It is known that a voter votes for the incumbent if and only if he desires the incumbent's first-term policy to be implemented in the next period. Using (2), this means that a type-μ voter votes for the incumbent when $x = 1$ if and only if

$$(4)\quad \rho y + (1-\rho)\mu \geq 1 - \mu$$

$$\rightarrow y \geq \underline{y}_1(\mu) \equiv \mu + (2/\rho)[(1/2) - \mu].$$

When instead $x = 2$, a type-μ voter votes for the incumbent if and only if

$$(5)\quad \rho y + (1-\rho)(1-\mu) \geq \mu$$

$$\rightarrow y \geq \underline{y}_2(\mu) \equiv (1-\mu) + (2/\rho)[\mu - (1/2)].$$

Using (4)–(5) and letting $x \in \{1, 2\}$ denote the incumbent's first-term policy, the optimal voting rule is

$$(6)\quad t^*(\mu, x, y) = \begin{cases} I & \text{if } y \geq \underline{y}_x(\mu) \\ C & \text{otherwise.} \end{cases}$$

According to this voting rule, a voter votes for the incumbent if and only if income (i.e., performance) is sufficiently great. Since $\underline{y}_1(\mu)$ is generally different from $\underline{y}_2(\mu)$, the standard to which an incumbent is held depends on the policy implemented (as well as on a voter's prior beliefs). Thus, both economic policy and the resulting level of economic activity influence voting behavior. The reason is as follows. A voter votes for

the incumbent if and only if he expects future income to be higher when the incumbent remains in office. A voter's beliefs on future income depend on his beliefs as to the policy that will be implemented and how effective that policy will be. The incumbent's past policy is instrumental in that it provides information as to the incumbent's future policy intentions while past performance provides information as to the effectiveness of the incumbent's policy. As a result, policy and performance jointly determine voting behavior in that a voter associates the incumbent with the policy he implemented and thereby votes for the incumbent if and only if that policy performed sufficiently well.

A noteworthy property of this voting rule is that a voter may vote for the incumbent even though he was formerly against the incumbent's policy. For example, if $\mu > \frac{1}{2}$, then a voter initially believes policy 1 is best but would vote for an incumbent who implemented policy 2 as long as $y \geq \underline{y}_2(\mu)$. In response to exceptional performance by a policy, a voter will modify his policy preferences. Note, however, that if a voter initially prefers, say, policy 1 then it takes greater income for him to vote for the incumbent when the incumbent implemented policy 2 than when he implemented policy 1. Given the voter's prior beliefs, he is more likely to attribute bad performance to a bad shock than to a bad policy (i.e., a low value of ω_x) when policy 1 is implemented as opposed to when policy 2 is implemented.

Finally, one can use this voting rule to provide insight as to when voting behavior is relatively sensitive to policy. For this purpose consider the difference in the reelection criteria:

$$(7)\quad \left|\underline{y}_1(\mu) - \underline{y}_2(\mu)\right| = 2\left|(1/2) - \mu\right|\left[1 + 2\left(\sigma_\varepsilon^2/\sigma^2\right)\right].$$

When $\underline{y}_1(\mu) = \underline{y}_2(\mu)$, a voter requires the same level of income to vote for the incumbent irrespective of the policy implemented. Hence, voting depends only on performance. As $|\underline{y}_1(\mu) - \underline{y}_2(\mu)|$ increases, the choice of policy has a bigger impact on the standard applied, so that voting behavior becomes more sensitive to policy vis-à-vis performance. According to (7), the more uncertain are voters as to which policy is best (the higher is σ^2) and the larger the impact of policy on income (the lower is σ_ε^2), the more sensitive electoral outcomes are to performance. In both cases, performance is a more accurate signal of the effectiveness of the incumbent's policy. Since voters believe that the incumbent is relatively likely to implement the same policy if reelected, past performance is then a more informative signal of future performance. Note that as voters become totally confident in their beliefs about which policy is best ($\sigma^2 \to 0$), $|\underline{y}_1(\mu) - \underline{y}_2(\mu)| \to \infty$ so that voting becomes purely policy-based (Rogoff, 1990). Alternatively, if voters are initially indifferent as to which policy is best ($\mu = \frac{1}{2}$), voting is purely performance-based (Alesina and Cukierman, 1990).

III. Equilibrium Policy Bias

Suppose each voter uses the voting rule in (6) and, given x and y, the median voter prefers the incumbent. If $x = 1$ then, since $\underline{y}_1(\mu)$ is decreasing in μ, all those voters whose type is greater than or equal to the median voter's type will also vote for the incumbent. Hence, the incumbent wins. If $x = 2$, then, since $\underline{y}_2(\mu)$ is increasing in μ, all voters whose type is less than or equal to the median voter's type will vote for the incumbent, so that the incumbent wins. It is then true that a candidate wins if and only if he is preferred by the median voter. From this point onward, μ will denote the median voter's type.

When one says that a strategy profile is a perfect Bayesian equilibrium, it is meant that there exists a set of beliefs such that the strategy profile and that set of beliefs form a perfect Bayesian equilibrium. Theorem 1 shows that (1) and (6) comprise an equilibrium when the incumbent is sufficiently policy-motivated; that is, the marginal utility of income, θ, is large relative to the incremental value to holding office, k.

THEOREM 1: *If and only if θ/k is sufficiently large does there exist $\alpha^* \in (0,1)$ such that (1) and (6) is a perfect Bayesian equilibrium. If α^* exists, then it is unique.*

PROOF:

Since (6) has been shown to be an optimal strategy $\forall\ \mu \in [0,1]$, it is only necessary to show that (1) is a best response to $t^*(\cdot)$ $\forall\ \alpha \in [0,1]$. To begin, let $\lambda(\alpha, x)$ denote the probability that a type-α incumbent assigns to being reelected, given that he implements policy x. In deriving $\lambda(\cdot)$, first note that, if he implements policy 1, reelection occurs if and only if the median voter prefers policy 1, which occurs if and only if $\rho y + (1-\rho)\mu \geq 1-\mu$. Substituting $\alpha+\varepsilon$ for y (since a type-α incumbent believes $y = \alpha+\varepsilon$) and solving for ε, a type-α incumbent expects to be reelected if and only if $\varepsilon \geq (1/\rho)[1-(2-\rho)\mu]-\alpha$. Letting $H(\cdot)$ denote the cumulative distribution function of ε, it follows that

$$(8)\quad \lambda(\alpha,1) = \int_0^1 \{1 - H([1/\rho][1-(2-\rho)\mu]-\alpha)\}G'(\mu)\,d\mu.$$

Therefore, a type-α incumbent's expected utility from implementing policy 1 in his first term is

$$(9)\quad E[U|\alpha,1] = (\theta\alpha+k) + \delta\{\lambda(\alpha,1)[\theta\max\{\alpha,1-\alpha\}+k] + [1-\lambda(\alpha,1)] \times [f\theta\alpha + (1-f)\theta(1-\alpha)]\}$$

where f is the probability that the challenger implements policy 1. It is straightforward to show that $E[U|\alpha,1]$ is strictly increasing in α since higher α increases expected income which is directly valued when $\theta > 0$ and also raises the probability of reelection. Similarly, one can derive

$$(10)\quad E[U|\alpha,2] = \theta(1-\alpha)+k + \delta\{\lambda(\alpha,2)[\theta\max\{\alpha,1-\alpha\}+k] + [1-\lambda(\alpha,2)] \times [f\theta\alpha + (1-f)\theta(1-\alpha)]\}$$

where

$$\lambda(\alpha,2) = \int_0^1 \{1 - H([1/\rho][(2-\rho)\mu-1]+\alpha)\}G'(\mu)\,d\mu.$$

$E[U|\alpha,2]$ can be shown to be strictly decreasing in α.

Define

$$\psi(\alpha) \equiv E[U|\alpha,1] - E[U|\alpha,2] = \theta(2\alpha-1) + \delta\Delta(\alpha)[\theta\max\{\alpha,1-\alpha\} + k - (f\theta\alpha + [1-f]\theta[1-\alpha])]$$

where $\Delta(\alpha) \equiv \lambda(\alpha,1) - \lambda(\alpha,2)$. It follows that (1) is a best response if and only if $\psi(\alpha) < 0\ \forall\ \alpha \in [0,\alpha^*]$ and $\psi(\alpha) > 0\ \forall\ \alpha \in (\alpha^*,1]$. Since $\psi(\cdot)$ is continuous and increasing in α, (1) is then a best response if and only if $\psi(\alpha^*) = 0$. Next note that there exists $\alpha^* \in (0,1)$ such that $\psi(\alpha^*) = 0$ if and only if $\psi(0) < 0 < \psi(1)$. It is straightforward to show that $\psi(0) < 0 < \psi(1)$ if and only if

$$(11)\quad \theta/k > \kappa \equiv \max\{\delta\Delta(0)/[1-\delta f\Delta(0)], -\delta\Delta(1)/[1+\delta(1-f)\Delta(1)]\}.$$

This proves Theorem 1. It is important to note that it is possible for $\Delta(1) > 0 > \Delta(0)$, in which case $\kappa < 0$. In that event, $\theta/k > \kappa$ is automatically satisfied since $\theta \geq 0$ and $k > 0$.

Given this equilibrium, the first task is to assess the policy bias created by the incumbent's desire to improve his chances of reelection. Toward this end, it will prove to be useful to have a measure of the anticipated popularity of a policy. Recall that $\Delta(\alpha)$ is the difference between the probabilities of reelection from implementing policy 1 and policy 2. When $\Delta(\alpha) > 0$, the advantage from policy 1 could be due to policy 1 being more effective (which is believed to be true when $\alpha > \frac{1}{2}$) or due to the fact that it is relatively likely that the median voter prefers policy 1 [roughly, $G(\cdot)$ puts relatively more weight

over $(\frac{1}{2},1]$]. In order to control for the former effect, I propose to use $\Delta(\frac{1}{2})$ as a measure of policy popularity among voters, since both policies are believed to be equally effective. I say that policy 1 is more popular than policy 2 if and only if $\Delta(\frac{1}{2})>0$.

Proposition 1 shows that some incumbent types do not implement the policy believed to be most effective and instead implement the policy that is expected to be more popular.

PROPOSITION 1: *If* $\Delta(\frac{1}{2})\gtreqless 0$, *then* $\alpha^*\lesseqgtr\frac{1}{2}$.

PROOF:

First note that $\psi(\frac{1}{2})=k\Delta(\frac{1}{2})$ so that $\psi(\frac{1}{2})\gtreqless 0$ as $\Delta(\frac{1}{2})\gtreqless 0$. If $\Delta(\frac{1}{2})>0$, then $\psi(\frac{1}{2})>0$. Since $\psi(\cdot)$ is increasing in α and $\psi(\alpha^*)=0$, then $\alpha^*<\frac{1}{2}$. Hence, if $\Delta(\frac{1}{2})>0$, then $\alpha^*<\frac{1}{2}$. By analogous means, one can complete the proof of Proposition 1.

Suppose $\Delta(\frac{1}{2})>0$ so that prior information suggests that voters are more likely to be *ex ante* in favor of policy 1. There are two factors at play which determine the incumbent's optimal policy. Ceteris paribus, reelection is more likely when the more popular policy is implemented as the median voter applies a less stringent standard in voting for the incumbent. On the other hand, implementing the policy that one believes is more effective generates higher expected income which the incumbent directly values (when $\theta>0$) and which results in a higher probability of achieving a given income criterion for reelection. There is no trade-off when the policy that the incumbent believes is more effective is also the more popular one. Thus one finds that $x^*(\alpha)=1$ for all $\alpha\in(\frac{1}{2},1]$. When α is less than but close to $\frac{1}{2}$, the incumbent believes that both policies generate roughly the same expected income. However, since policy 1 is more popular, the incumbent's probability of reelection is strictly higher from implementing it. As a result, these incumbent types implement policy 1 even though they find it to be less effective than policy 2. When α is sufficiently small, the income differential between the two policies is sufficiently great that the incumbent prefers to implement policy 2 despite its anticipated unpopularity. This occurs for $\alpha<\alpha^*$. Strategic manipulation of policy then results in incumbent types over $[\alpha^*,\frac{1}{2})$ implementing the more popular policy even though it is expected to generate lower income.

By the preceding analysis, the more popular policy will be implemented by all incumbent types who believe it to be the most effective policy (and thus will implement it again if reelected) but also by some incumbent types who believe it to be the weaker policy (and thus will not implement it again if reelected). As a result, the more popular policy is a noisy signal of the incumbent's future policy intentions. In contrast, the less popular policy is a perfect signal of the incumbent's future policy intentions, since he implements it only if he believes it is the most effective policy. Hence, he is certain to implement it again if reelected.

Though the more popular policy is a noisy signal, it is important to emphasize that the incumbent's first-term policy always contains information as to his second-term policy as long as the incumbent is sufficiently policy-motivated [i.e., $\theta/k>\kappa$, where κ is defined in (11)]. This restriction on the degree of policy motivation is trivially satisfied when the incumbent's prior beliefs on the median voter's type, $G(\cdot)$, are symmetric as $\kappa<0$. Since then $\Delta(\frac{1}{2})=0$ (so that both policies are equally popular), it follows from Proposition 1 that $\alpha^*=\frac{1}{2}$. In this special case, both policies are perfect signals of the incumbent's future policy intentions. By a continuity argument, Proposition 2 establishes that if neither policy is believed to be heavily favored by the electorate then the incumbent's first-term policy will contain information about his beliefs and thereby about the policy he would implement if reelected. Since the proof is straightforward, it is omitted.

PROPOSITION 2: *Assume that* $G(\cdot)$ *results in* $\Delta(\frac{1}{2})=0$ *and* $\{G^\tau(\cdot)\}$ *converges weakly to* $G(\cdot)$. *Then* $\lim_{\tau\to\infty}\alpha^*(G^\tau(\cdot))=\frac{1}{2}$.

It follows from Propositions 1 and 2 that if policies are not too disparate in terms of

their anticipated popularity, an incumbent's first-term policy always contains information as to his policy if reelected. This result holds even if $\theta = 0$ so that politicians are purely office-motivated. In contrast to preceding analyses which investigate the informativeness of an incumbent's policy, information can be conveyed even when the incumbent is purely office-motivated.[7] Such an incumbent will implement the less popular policy when he believes that it is sufficiently effective, because exceptional performance can induce voters to change their policy preferences. This result highlights the role of voters' policy preferences being endogenous. If instead voters' policy preferences are fixed, then a purely office-motivated incumbent would always choose the more popular policy so that his policy choice would contain no information as to his future policy intentions.

While I have shown that there exists an equilibrium in which the incumbent's policy contains information about his future policy intentions, there do exist other perfect Bayesian equilibria. If the incumbent is sufficiently office-motivated, there exist both pooling equilibria—so that the incumbent implements the same policy regardless of his type—and nonpooling equilibria in which an incumbent implements a policy if and only if he believes it to be sufficiently ineffective (contrary to the equilibrium of Theorem 1).[8] However, Theorem 2 shows that if the incumbent is sufficiently policy-motivated then (1) and (6) form the unique perfect Bayesian equilibrium.

THEOREM 2: *(1) and (6) is the unique perfect Bayesian equilibrium if and only if*

$$\theta/k > \max\{\delta/[1-\delta(1-f)], \delta/(1-\delta f)\}.$$

(The proof is available from the author upon request.)

IV. Comparative Statics

In this section, I perform comparative-static exercises with respect to the equilibrium defined in (1) and (6). To ensure existence, it is assumed that θ/k is sufficiently large according to Theorem 1. All proofs of comparative statics are in the Appendix.

A. *The Effect of Policy Popularity*

In order to improve their chances of reelection, some incumbent types will implement the policy that is likely to be favored by the median voter even though it does not maximize expected income. Proposition 3 shows that this effect becomes more pronounced the more popular a policy is expected to be. Note that if $G^0(\cdot)$ stochastically dominates $G^{00}(\cdot)$ in the first degree then there is a higher probability that the median voter prefers policy 1 and, by Proposition 3, a higher probability that the incumbent implements policy 1.

PROPOSITION 3: *If $G^0(\cdot)$ stochastically dominates $G^{00}(\cdot)$ in the first degree, then $\alpha^*(G^0(\cdot)) < \alpha^*(G^{00}(\cdot))$.*

This theory then predicts that a policy is more likely to be implemented, the stronger is support for it as revealed through pre-election polls. While this result is unsurprising, I am unaware of it having been formally derived as an implication of preceding models.

B. *The Effect of Policy Motivation*

A critical property of equilibrium is that some incumbent types hide their true policy intentions by implementing the more popular policy. This has the implication of making the more popular policy an imperfect signal of the policy the incumbent will implement if reelected. A measure of this ambiguity is the "number" of incumbent types who strategically manipulate policy: $|\alpha^* - \frac{1}{2}|$. This is a meaningful measure because a marginal rise in $|\alpha^* - \frac{1}{2}|$ results in a

[7] A possible exception is Austen-Smith and Banks (1989).

[8] The nonpooling equilibrium mentioned has $x(\alpha) = 1 \ \forall \ \alpha \in [0, \alpha^0]$ and $x(\alpha) = 2 \ \forall \ \alpha \in (\alpha^0, 1]$ for some $\alpha^0 \in (0,1)$. Since all actions are implemented in equilibrium with positive probability, this equilibrium will survive any known refinement. One can also show that pooling equilibria survive the intuitive criterion (In-Koo Cho and David M. Kreps, 1987) when candidates are sufficiently office-motivated; that is, θ is sufficiently close to 0.

lower probability that the incumbent will implement the same policy if reelected. Since a voter votes for the incumbent in the hope that he will implement the same policy in his second term, this lower probability is detrimental to voters. The ensuing analysis assesses how changes in parameters influence the degree of ambiguity.

As the incumbent becomes more policy-motivated, he cares more about the income generated during his time in office. As a result, an incumbent will be less inclined to implement the more popular policy when he expects it to perform relatively poorly. Since then more incumbent types implement the policy they believe is more effective, an incumbent's policy is a better signal of his future policy intentions, as stated in Proposition 4.[9]

PROPOSITION 4: *If* $\alpha^* \neq \frac{1}{2}$, *then* $\partial|\alpha^* - \frac{1}{2}|/\partial\theta < 0$.

C. *The Effect of Voter Uncertainty*

A singular feature of the present model is that voters are uncertain about which policy is best. This has played a central role in my analysis as it results in voters' policy preferences being endogenous to performance. It is then of particular interest to consider how equilibrium behavior responds to a change in voter uncertainty.

When voters become less confident in their prior beliefs as to the efficacy of different policies (as represented by a rise in σ^2), they assign more weight to observed performance in updating these beliefs [see (2) and (3)]. As a result, their voting behavior becomes less dependent on their initial opinions and more sensitive to performance. This is reflected in the difference in the reelection criteria shrinking: $|y_1(\mu) - y_2(\mu)| = 2|\frac{1}{2} - \mu|[1 + 2(\sigma_\varepsilon^2/\sigma^2)]$ is decreasing in σ^2. The effect of this change in voters' standards on the incentives of the incumbent is as follows. Suppose $\alpha^* < \frac{1}{2}$ so that some incumbent types implement policy 1 even though they think it will produce less income. Recall that this is optimal because the reelection criterion depends on both which policy is chosen and the prior beliefs of the median voter. If it is sufficiently likely that the median voter is predisposed to policy 1, then an incumbent might choose policy 1 even if he believes it will produce lower income on average. As I argued above, the difference in reelection criteria shrinks as voters become more uncertain, so that the probability of reelection is less dependent on the identity of the median voter and more dependent on how well the policy performs. An incumbent is then less inclined to implement a policy simply because it is likely to be initially well received and more inclined to implement the policy that maximizes income. As a result, there are fewer incumbent types who strategically implement the more popular policy so that $|\alpha^* - \frac{1}{2}|$ shrinks. Since it is more likely that the policy an incumbent implemented is the one he would implement if reelected, an incumbent's policy becomes a more informative signal of his second-term policy.

While we believe that this result is general, it has only been proved for a class of distributions on the median voter's type. This class is defined below:

$$\Omega \equiv \{G(\cdot):[0,1]\to[0,1] \mid G'(\mu) \geq G'(1-\mu)\ \forall\ \mu \in [0,\tfrac{1}{2}] \text{ or } G'(\mu) \leq G'(1-\mu)\ \forall\ \mu \in [0,\tfrac{1}{2}]\}. \tag{12}$$

PROPOSITION 5: *If* $\alpha^* \neq \frac{1}{2}$ *and* $G(\cdot) \in \Omega$ *then* $\partial|\alpha^* - \frac{1}{2}|/\partial\sigma^2 < 0$.

Ω encompasses many distributions, including all unimodal distributions for which the density function is symmetric around its maximum and all distributions for which the density function is monotonic (though the latter is not a particularly likely set of beliefs on the median voter's type). Note that the uniform distribution, which is assumed

[9] All of the ensuing comparative statics are derived only when $\alpha^* \neq \frac{1}{2}$ because $\psi(\cdot)$ is not differentiable at $\alpha = \frac{1}{2}$.

by Alesina and Cukierman (1990), lies in Ω. Finally, it is important to emphasize that the requirement that $G(\cdot)$ lie in Ω is a sufficient condition. For example, one can show that $\partial|\alpha^* - \frac{1}{2}|/\sigma^2 < 0$ when $G(\cdot)$ lies in an appropriately defined neighborhood of the set Ω.[10]

V. Empirical Implications

As reviewed at the beginning of the paper, previous theories on the economic determinants of electoral outcomes stress the dominant role of an incumbent's policies in determining the success of his bid for reelection. Performance plays a role only as a noisy signal of unobservable policies. In contrast, the theory developed here predicts that measures of performance are influential in elections even when policies are observed by voters. It is predicted that the probability of reelection is increasing in performance (e.g., per capita real GNP) and increasing in the popularity of a policy (measured prior to its implementation).

There are several ways to test these competing theories. One way is to estimate the probability of reelection as a function of measures of performance and a policy-popularity index. This index could be a weighted average of approval ratings for particular policies (derived from public opinion polls). If it is not possible to construct such an index, one could substitute the policy instruments themselves. In contrast to preceding theories, my theory predicts that measures of performance will be significantly correlated with electoral outcomes even when the incumbent's policies are controlled for in the empirical specification.[11]

To the best of my knowledge, there has been no empirical work that simultaneously controls for both policy and performance in assessing the determinants of electoral outcomes. However, there is considerable empirical evidence to substantiate the claim that the performance of the economy is correlated with voting (see the reviews in Kirsten Renwick Monroe [1984] and Kiewiet and Rivers [1985]). There has been considerably less empirical work that explores the impact of policies. However, what work there is finds that policies are influential. Henry W. Chappell, Jr., and William R. Keech (1985) show that monetary policy influences presidential popularity, while Michael S. Lewis-Beck (1988) provides evidence from public opinion surveys that voters appear to take account of policies in deciding how to vote.

While one could test the present theory using presidential elections, an alternative data set could be compiled from mayoral elections. The president's control over policy instruments can be problematic when he faces a nonsupportive Congress, while the presence of other players in the policy arena is less of a concern for a city mayor. In this setting, voter uncertainty can be interpreted as uncertainty as to how taxes and expenditures translate into municipal services. For example, voters may be uncertain (and differ in their beliefs) as to the effect of a change in police expenditure on the crime rate or the value of an increase in municipal services resulting from a rise in property taxes. Policy instruments could include property tax rates and expenditures on police and education. The associated measures of performance could include property values (a measure of the value of services to property owners), the crime rate, and the high-school graduation rate. The plan would be to estimate the probability that a mayor is reelected as a function of these policy instruments and measures of performance. One would also need to control for observable variables outside of the control of the mayor that would affect these measures of performance. For example, the general state of the economy affects the crime rate and housing prices.

[10] It has also been proved that $\partial|\alpha^* - \frac{1}{2}|/\partial\sigma^2 < 0$ when σ^2 is sufficiently small without any restrictions on $G(\cdot)$.

[11] Realistically, this will be a difficult hypothesis to test, due to omitted-variables bias. Some relevant policy instruments will inevitably be left out of the empirical specification. If the included measures of performance are affected by these omitted policy instruments, one could get a significant coefficient on performance even if the theory is false.

In addition to making predictions about the determinants of electoral outcomes, my theory also predicts the determinants of an elected official's policies. As shown in the preceding section, a policy is more likely to be chosen, the more popular it is. Furthermore, the policy that is more popular is less likely to be chosen, the more uncertain are voters. In translating these comparative statics into testable hypotheses, define p as the proportion of people polled in favor of policy 1. Assuming that p is a noisy estimate of the median voter, let $G(\cdot\,|p)$ be the incumbent's beliefs on the median voter's type conditional on the public opinion poll. If one supposes that $p' > p''$ implies that $G(\cdot\,|p')$ stochastically dominates $G(\cdot\,|p'')$ in the first degree, it follows from Proposition 3 that the probability that a policy is implemented is increasing in the prior approval rating for that policy.[12] Benjamin I. Page and Robert Y. Shapiro (1983) provide evidence supportive of this hypothesis in that they find that changes in public opinion cause changes in policies. Proposition 5 states that the more popular policy (say, the policy with $p > \frac{1}{2}$) is less likely to be implemented the greater is voter uncertainty. If one assumes voter uncertainty over which policy is best to be greater for new policy issues, then the theory would predict that the probability that a policy is implemented is greater for a new issue when $p < \frac{1}{2}$ and lower when $p > \frac{1}{2}$.

VI. Concluding Remarks

An essential feature of the electoral process is that voters face considerable uncertainty when deciding for whom to vote. Due to the central role of voter uncertainty, there has been a growing amount of research which explores the incentives of elected officials to exploit this uncertainty in order to improve their chances of reelection. Generally, voters have been modeled as being uncertain of politicians' preferences or their levels of competency. A new source of voter uncertainty was introduced in this paper. Voters were assumed to be uncertain as to the efficacy of different policies. By being able to learn over time, voters' policy preferences were found to be endogenous to the level of income. Naturally, voters (as well as politicians) were allowed to have different opinions as to the best policy.

My analysis generated several new findings. In contrast to preceding work, I found that electoral outcomes are influenced by both an incumbent politician's past policies and the level of economic activity during his time in office. Performance has a bigger impact on the electoral outcome, the more uncertain are voters as to which policy is more productive. Whether policy or performance tends to be the dominant factor in elections was shown to have important implications for the behavior of elected officials. The less sensitive are electoral outcomes to performance relative to policy, the greater is the manipulation of policy for the purpose of reelection. Specifically, an incumbent is more likely to implement the policy that he believes will be well received rather than the policy that he believes will maximize expected income. This causes an incumbent's past policies to be a less informative signal of his future policy intentions. By manipulating policy in this manner, an incumbent improves his chances of reelection and thereby affects future policy as well as current policy.

Appendix

PROOF OF PROPOSITION 3:

G^0 stochastically dominates G^{00} in the *first degree* if $G^0(\mu) \leq G^{00}(\mu)\ \forall\ \mu \in [0,1]$ and $G^0(\mu) < G^{00}(\mu)$ for some $\mu \in [0,1]$. First note that G enters $E[U|\alpha,1]$ only through the probability of reelection [see (9)]. Since the integrand in (8) is strictly increasing in μ then $\lambda(\alpha,1)$ is strictly higher under G^0 than under G^{00}. Hence, $E[U|\alpha,1]$ is strictly higher under G^0 than under G^{00}.

[12] Selection bias could be a problem here. If it is known that voters are predisposed to a particular policy, then a candidate whose track record tends to support that policy is more likely to be reelected. Assuming voters' preferences are correlated over time, one could find a correlation between current support for a policy and the probability that an elected official would implement it even if elected officials always implement the policy they believe is best.

Turning to $\lambda(\alpha,2)$ in (10), the integrand is strictly decreasing in μ so that $\lambda(\alpha,2)$ is strictly lower under G^0 than under G^{00}. Hence, $E[U|\alpha,2]$ is strictly lower under G^0 than under G^{00}.

It follows that $\psi(\alpha;G^0)>\psi(\alpha;G^{00})\ \forall\ \alpha\in[0,1]$. Since α^* is defined by $\psi(\alpha^*;G)=0$ then $\psi(\alpha^*(G^{00});G^0)>\psi(\alpha^*(G^{00});G^{00})=0$. Given that $\psi(\cdot)$ is increasing in α, it is true that $\alpha^*(G^0)<\alpha^*(G^{00})$.

PROOF OF PROPOSITION 4:

Recall that α^* is defined by $\psi(\alpha^*)=0$. Given that $\partial\psi(\cdot)/\partial\theta\neq 0$ and $\psi(\cdot)$ is differentiable in θ and $\alpha\ \forall\ \alpha\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, then

$$\text{(A1)}\quad \frac{\partial\alpha^*}{\partial\theta}=\frac{-\left[\partial\psi(\alpha^*(\theta),\theta)/\partial\theta\right]}{\left[\partial\psi(\alpha^*(\theta),\theta)/\partial\alpha\right]}.$$

Since $\partial\psi(\cdot)/\partial\alpha>0\ \forall\ \alpha\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, it follows that

$$\begin{aligned}\text{(A2)}\quad &\operatorname{sign}\{\partial\alpha^*/\partial\theta\}\\ &=\operatorname{sign}\{-\partial\psi(\alpha^*(\theta),\theta)/\partial\theta\}\\ &=\operatorname{sign}\{2[\alpha^*-\tfrac{1}{2}]\\ &\qquad+\delta\Delta(\alpha^*)[\max\{\alpha^*,1-\alpha^*\}\\ &\qquad-f\alpha-(1-f)(1-\alpha^*)]\}.\end{aligned}$$

When $\alpha^*\in(0,\frac{1}{2})$, the right-hand side of (A2) is $\operatorname{sign}\{2(\alpha^*-\frac{1}{2})[1-\delta f\Delta(\alpha^*)]\}$, which is negative as $\delta\leq 1$, $f<1$, and $\Delta(\alpha^*)\leq 1$. Given that $\alpha^*<\frac{1}{2}$ and $\partial\alpha^*/\partial\theta<0$, it follows that $\partial|\alpha^*-\frac{1}{2}|/\partial\theta<0$. When instead $\alpha^*\in(\frac{1}{2},1)$, then the right-hand side of (A2) is $\operatorname{sign}\{2(\alpha^*-\frac{1}{2})[1+\delta(1-f)\Delta(\alpha^*)]\}$, which is positive. Hence, $\partial|\alpha^*-\frac{1}{2}|/\partial\theta<0$.

PROOF OF PROPOSITION 5:

Since α^* is defined by $\psi(\alpha^*)=0$, then

$$\text{(A3)}\quad \frac{\partial\alpha^*}{\partial\sigma^2}=\frac{-\left[(\partial\psi(\alpha^*,\rho)/\partial\rho)(\partial\rho/\partial\sigma^2)\right]}{\left[\partial\psi(\alpha^*,\rho)/\partial\alpha\right]}.$$

Given that $\partial\rho/\partial\sigma^2>0$, then

$$\operatorname{sign}\{\partial\alpha^*/\partial\sigma^2\}=\operatorname{sign}\{-\partial\psi(\alpha^*,\rho)/\partial\rho\}.$$

Since ρ enters $\psi(\cdot)$ only through $\Delta(\cdot)$ and given $\psi(\cdot)$ is increasing in Δ, then

$$\begin{aligned}&\operatorname{sign}\{-\partial\psi(\alpha^*,\rho)/\partial\rho\}\\ &\quad=\operatorname{sign}\{-\partial\Delta(\alpha^*,\rho)/\partial\rho\}.\end{aligned}$$

To prove Proposition 5, I will show that when $\alpha^*\neq\frac{1}{2}$ and $G(\cdot)\in\Omega$ then $\partial\Delta(\alpha^*(\rho),\rho)/\partial\rho\gtrless 0$ as $\alpha^*\gtrless\frac{1}{2}$.

Suppose $G'(1-\mu)\geq G'(\mu)\ \forall\ \mu\in[0,\frac{1}{2}]$ and $G'(1-\mu)>G'(\mu)$ for some $\mu\in[0,\frac{1}{2})$. I begin by proving that this implies $\alpha^*<\frac{1}{2}$. Define $\hat{G}(\cdot)$ by $\hat{G}'(\mu)=G'(\mu)+\beta(\mu)$ and $\hat{G}'(1-\mu)=G'(1-\mu)-\beta(\mu)$ where $\beta(\mu)\equiv[G'(1-\mu)-G'(\mu)]/2,\ \mu\in[0,\frac{1}{2})$. Since $\hat{G}'(\mu)=\hat{G}'(1-\mu)\ \forall\ \mu\in[0,\frac{1}{2}]$ then $\hat{G}(\cdot)$ is symmetric around $\frac{1}{2}$ which implies that $\alpha^*(\hat{G})=\frac{1}{2}$. Given that $G'(\mu)\lessgtr\hat{G}'(\mu)$ as $\mu\lessgtr\frac{1}{2}$, then the single-crossing property is satisfied, which means that $G(\cdot)$ stochastically dominates $\hat{G}(\cdot)$ in the first-degree. From Proposition 3, it follows that $\alpha^*(G)<\alpha^*(\hat{G})$ or $\alpha^*<\frac{1}{2}$.

Given that $\alpha^*<\frac{1}{2}$ then $\partial|\alpha^*-\frac{1}{2}|/\partial\sigma^2<0$ if and only if $\partial\Delta(\alpha^*,\rho)/\partial\rho<0$. It is straightforward to derive equation (A4), below. Since $1-2\mu\geq 0$ as $\mu\leq\frac{1}{2}$, then a sufficient condition for (A4) to be negative is

---

$$\begin{aligned}\text{(A4)}\quad &\partial\Delta(\alpha^*,\rho)/\partial\rho\\ &=(2/\rho^2)\int_0^{\frac{1}{2}}(1-2\mu)\{H'([1/\rho][1-(2-\rho)\mu]-\alpha^*)G'(\mu)-H'([1/\rho][1-(2-\rho)(1-\mu)]-\alpha^*)G'(1-\mu)\}\,d\mu\end{aligned}$$

that the integrand is negative $\forall\ \mu \in [0, \frac{1}{2})$. One concludes that

(A5) if $H'([1/\rho][1-(2-\rho)(1-\mu)]-\alpha^*)G'(1-\mu)$
$> H'([1/\rho][1-(2-\rho)\mu]-\alpha^*)G'(\mu)$
$\forall\ \mu \in [0, \frac{1}{2})$
then $\partial\Delta(\alpha^*, \rho)/\partial\rho < 0$.

It is straightforward to show that

(A6) $H'([1/\rho][1-(2-\rho)(1-\mu)]-\alpha^*)$
$> H'([1/\rho][1-(2-\rho)\mu]-\alpha^*)$
$\forall\ \mu \in [0, \frac{1}{2})$.

It follows that a sufficient condition for (A5) to hold is that $G'(1-\mu) \geq G'(\mu)\ \forall\ \mu \in [0, \frac{1}{2}]$ which is true by supposition. I have then shown that if $\alpha^* < \frac{1}{2}$ then $\partial\Delta(\alpha^*, \rho)/\partial\rho < 0$. This proves that $\partial|\alpha^* - \frac{1}{2}|/\partial\sigma^2 < 0$.

When $G'(\mu) \geq G'(1-\mu)\ \forall\ \mu \in [0, \frac{1}{2}]$ and $G'(\mu) > G'(1-\mu)$ for some $\mu \in [0, \frac{1}{2})$ then $\alpha^* > \frac{1}{2}$, and the proof is analogous. When $G'(\mu) = G'(1-\mu)\ \forall\ \mu \in [0, \frac{1}{2}]$ then $\alpha^* = \frac{1}{2}$, which violates the condition of Proposition 5 that $\alpha^* \neq \frac{1}{2}$.

REFERENCES

Alesina, Alberto and Cukierman, Alex, "The Politics of Ambiguity," *Quarterly Journal of Economics*, November 1990, *105*, 829–50.

Aumann, Robert J., "Agreeing to Disagree," *Annals of Statistics*, November 1976, *4*, 1236–9.

Austen-Smith, David and Banks, Jeffrey S., "Electoral Accountability and Incumbency," in Peter C. Ordeshook, ed., *Strategic Models of Politics*, Ann Arbor: University of Michigan Press, 1989, pp. 121–48.

_____ **and Riker, William,** "Asymmetric Information and the Coherence of Legislation," *American Political Science Review*, September 1987, *81*, 897–918.

Banks, Jeffrey S. and Sundaram, Rangarajan K., "Incumbents, Challengers, and Bandits: Bayesian Learning in a Dynamic Choice Model," mimeo, University of Rochester, August 1990.

Bernhardt, M. Daniel and Ingberman, Daniel E., "Candidate Reputations and the 'Incumbency Effect'," *Journal of Public Economics*, June 1985, *27*, 47–67.

Chappell, Henry W., Jr. and Keech, William R., "The Political Viability of Rule-Based Monetary Policy," *Public Choice*, 1985, *46* (2), 125–40.

Cho, In-Koo and Kreps, David M., "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, May 1987, *102*, 179–221.

Cukierman, Alex and Meltzer, Allan H., "A Positive Theory of Discretionary Policy, the Costs of Democratic Government, and the Benefits of a Constitution," *Economic Inquiry*, July 1986, *24*, 367–88.

DeGroot, Morris H., *Optimal Statistical Decisions*, New York: McGraw-Hill, 1970.

Downs, Anthony, *An Economic Theory of Democracy*, New York: Harper and Row, 1957.

Ferejohn, John A., "Incumbent Performance and Electoral Control," *Public Choice*, 1986, *50* (1–3), 5–25.

Fiorina, Morris P., *Retrospective Voting in American National Elections*, New Haven: Yale University Press, 1981.

Frey, Bruno S., Pommerehne, Werner W., Schnedier, Friedrich and Gilbert, Guy, "Consensus and Dissension Among Economists: An Empirical Inquiry," *American Economic Review*, December 1984, *74*, 986–94.

Gilligan, Thomas W. and Krehbiel, Keith, "Asymmetric Information and Legislative Rules with a Heterogeneous Committee," *American Journal of Political Science*, May 1989, *33*, 459–90.

Hess, Gregory D., "Voting and the Intertemporal Selection of Tax Rates in a Macro-Economy," *Economics and Politics*, March 1991, *3*, 41–62.

Hibbs, Douglas A., Jr., *The American Political Economy: Macroeconomics and Electoral Politics*, Cambridge, MA: Harvard University Press, 1987.

Hintikka, Jakkoo, "Impossible Possible Worlds Vindicated," *Journal of Philo-*

sophical Logic, November 1975, *4*, 475–84.

Key, V. O., Jr., *The Responsible Electorate*, Cambridge, MA: Belknap, 1966.

Kiewiet, D. Roderick and Rivers, Douglas, "A Retrospective on Retrospective Voting," in Heinz Eulau and Michael S. Lewis-Beck, eds., *Economic Conditions and Electoral Outcomes: The United States and Western Europe*, New York: Agathon, 1985, pp. 207–31.

Lewis-Beck, Michael S., *Economics and Elections*, Ann Arbor: University of Michigan Press, 1988.

Lipman, Barton L., "How to Decide How to Decide How to...: Modeling Limited Rationality," *Econometrica*, September 1991, *59*, 1105–1125.

Lott, John R., Jr. and Reed, W. Robert, "Shirking and Sorting in a Political Market with Finite-Lived Politicians," *Public Choice*, April 1989, *61*, 75–96.

MacRae, C. Duncan, "A Political Model of the Business Cycle," *Journal of Political Economy*, April 1977, *85*, 239–63.

Monroe, Kristen Renwick, *Presidential Popularity and the Economy*, New York: Praeger, 1984.

Nordhaus, William D., "The Political Business Cycle," *Review of Economic Studies*, April 1975, *42*, 169–90.

Page, Benjamin I. and Shapiro, Robert Y., "Effects of Public Opinion on Policy," *American Political Science Review*, March 1983, *77*, 175–90.

Reed, W. Robert, (1989a) "Information in Political Markets: A Little Knowledge Can Be a Dangerous Thing," *Journal of Law, Economics, and Organization*, Fall 1989, *5*, 355–74.

_____, (1989b) "Retrospective and Prospective Voting," mimeo, Texas A&M University, September 1989.

Rogoff, Kenneth, "Equilibrium Political Budget Cycles," *American Economic Review*, March 1990, *80*, 21–36.

_____ **and Sibert, Anne,** "Elections and Macroeconomic Policy Cycles," *Review of Economic Studies*, January 1988, *55*, 1–16.

The Macroeconomics of Dr. Strangelove

By A. Andrew John, Rowena A. Pecchenino, and Stacey L. Schreft*

This paper examines the weapons-accumulation decisions of two adversarial countries in the context of a deterrence / conflict initiation game embedded in an overlapping-generations model. The demographic structure permits analysis of both within- and between-country intergenerational externalities caused by past weapons-accumulation decisions, as well as of intragenerational externalities from the adversary's current weapons accumulation. Zero accumulation is a possible equilibrium with both noncooperative and cooperative behavior. Countries may also accumulate weapons to the point where conflict initiation never occurs. Pareto-improving policies are generally available, but international cooperation need not be Pareto-improving. (*JEL* D62, D74, E21)

The Cold War has ended, and President George Bush has declared the beginning of a "new world order." This new world order has so far been characterized by the intensification of regional disputes, civil strife, and civil war; its implications for arms proliferation and arms control are still unclear. This paper considers both superpower conflict and the transition to international cooperation in the context of a model that emphasizes intergenerational linkages. Specifically, we explore the weapons-accumulation decisions of two adversarial countries in the context of a deterrence/conflict initiation game embedded in an overlapping-generations model. The demographic structure of the overlapping-generations model allows the analysis of both between- and within-country intergenerational external effects caused by past weapons-accumulation decisions, while the two-country setting permits consideration of international external effects caused by the adversary country's current weapons-accumulation decision. We analyze the noncooperative equilibrium and then consider the effects of a regime change characterized by international cooperation, but where the intergenerational externalities remain.

We emphasize intergenerational externalities in our model because they are intrinsically hard to internalize: those who bear their cost or receive their benefit are without representatives in the present generation. Thus their internalization requires intervention by long-lived government institutions. We also analyze the problem of such institutions.

The analysis draws on a number of diverse economic literatures. The model utilizes the overlapping-generations framework of Maurice Allais (1947) and Paul A. Samuelson (1958), since this demographic structure lends itself to analysis of situations where agents' actions have consequences beyond their own lifetimes. Positive intergenerational external effects play a role in models of endogenous growth (see e.g., Paul M. Romer, 1986; Nancy L. Stokey, 1988), while intergenerational diseconomies are the focus of Todd Sandler's (1982) work on intergenerational club goods. In this paper, as in work on environmental externalities (John et al., 1992), both positive and negative intergenerational effects interact. Finally, the model draws upon the game-theoretic arms-control literature (see e.g., Michael D. Intriligator and Dagobert L.

*John and Pecchenino: Department of Economics, Michigan State University, East Lansing, MI 48824. Schreft: Federal Reserve Bank of Richmond, Richmond, VA 23261. We thank David Schimmelpfennig for stimulating our thinking about intergenerational externalities, and Carl Davidson, John Goddeeris, Marvin Goodfriend, Daniel Hamermesh, Neil Wallace, and three anonymous referees for their helpful comments. The views expressed in this paper are solely those of the authors and do not reflect the views of the Federal Reserve Bank of Richmond or the Federal Reserve System. We take responsibility for all errors.

Brito, 1984; Steven J. Brams and D. Marc Kilgour, 1985; Frederick van der Ploeg and Aart J. de Zeeuw, 1990).

The main findings of the model are that equilibria in which neither country accumulates weapons are possible outcomes of both noncooperative and cooperative versions of the game and that countries may choose to accumulate weapons to the point where conflict initiation is so dangerous that it never occurs. The dynamic adjustment of the model to a steady state can be interpreted as an arms race: successive generations engage in individually rational but socially suboptimal arms accumulation. For second-best reasons, neither within-country intergenerational cooperation nor international cooperation necessarily leads to smaller weapons stocks or to Pareto improvements.

I. The Environment

A. *Preliminaries*

Consider an infinite-horizon world with two identical countries (called A and B). In each country, a new generation (called generation t) of agents is born at each date $t = 1, 2, \ldots$. There is no population growth, and each generation is normalized to be of size 1. Each member of generation t lives for one period with certainty, and for two periods with probability $\phi(\cdot)$. Further, each has preferences defined over consumption when young, $c^y(t)$, and when old, $c^o(t+1)$. These preferences are represented by a von Neumann-Morgenstern utility function:

$$(1) \quad U(c^y(t)) + \phi(\cdot)V(c^o(t+1))$$

where $\phi(\cdot)$ is derived below. Assume that $U(\cdot)$ and $V(\cdot)$ are both increasing, twice continuously differentiable, and strictly concave. Also assume that

$$\lim_{c^y \to 0} U'(\cdot) = \infty$$

$$\lim_{c^o \to 0} V'(\cdot) = \infty$$

$$V(0) \geq 0.$$

Each member of each generation $t \geq 1$ in each country is endowed with $e > 0$ units of the economy's single good when young and nothing when old. At date $t = 1$ each country collectively possesses a stock of $W(1) \geq 0$ weapons.

Any portion of the date-t endowment can be consumed at date t, yielding utility at t, or used to provide for future utility. To accomplish the latter, the consumption good can either be saved or used to add to the home economy's weapon stock. Each young agent has access to a storage technology, which is the sole vehicle for saving. Storage of $s(t)$ units of the good at date t yields $Rs(t)$ units of the good at date $t+1$, where $R > 0$ is the gross technological rate of return. Alternatively, units of the good can be converted into an equivalent number of units of weapons, $w(t)$. Accumulation of weapons at t adds to the stock of weapons that will be carried into date $t+1$, $W(t+1)$. The stock of weapons evolves according to

$$(2) \quad W(t+1) = (1-\delta)W(t) + w(t)$$

given the depreciation rate $\delta \in (0, 1]$ and $W(1)$. Weapons accumulation affects utility by altering the probability of survival, $\phi(\cdot)$, as discussed in the next subsection.

Since a primary focus of this paper is on external effects across generations, we abstract from the well-understood free-rider problems within a country and within a generation. We assume that the agents alive in each country at the beginning of each date elect a government for a one-period term. This government behaves myopically, carrying out policies made solely for the welfare of agents alive during its term in office. The government has the power to levy lump-sum taxes and transfers. Specifically, it levies lump-sum taxes on the young to achieve the desired stock of weapons. That is, an agent's choice of weapons can be interpreted as arising from the collective provision of national defense, a public good. This alloca-

tion could be achieved as a Lindahl equilibrium.[1]

B. *The Game Between Countries*

Each generation of young agents plays a one-shot game against its counterparts in the other country. At the start of the period, young agents choose the amount of their endowment that they wish to convert to weapons, taking as given the equivalent choice in the other country and the inherited stocks of weapons in both countries. From equation (2), these decisions give rise to the current weapons stocks (W^A, W^B) in the two countries. (For ease of presentation, we suppress the time index in this subsection.)

At the end of the period there is a "potential for conflict."[2] Specifically, there is a realization of a random variable that determines the probability that one side or the other will win, should either side decide to initiate a conflict.[3] A conflict has three characteristics. First, it has a winner and a loser; that is, there is a utility gain from winning and loss from losing. Second, we assume for simplicity that a conflict does not deplete either country's stock of weapons. Third, a conflict can escalate into a full-scale catastrophic war.[4]

Each side decides whether or not to initiate conflict. The gain from winning and loss from losing are proportional to utility when old and are the same no matter which side initiates. If the conflict does escalate, then there is mutual annihilation, giving zero utility in old age. The risk of this outcome generates the cost of initiating a conflict.

If either side initiates conflict, the probability of annihilation is $q = q(W^A, W^B)$, with $q_1 > 0$, $q_2 > 0$, $q_{11} < 0$, $q_1(x,x) = q_2(x,x)\ \forall\ x$ and $q(0,0) = 0$. The probability that country A wins a conflict is

$$p = \eta + f(W^A, W^B) \qquad \eta \sim U[0,1]$$

where η is a random variable determining the potential for conflict, and $f_1 > 0$, $f_2 < 0$, $f_{11} < 0$, $f(x,y) \equiv -f(y,x)\ \forall\ x, y$, and $f(x,x) \equiv 0\ \forall\ x$.[5] The probability that country B wins is $1 - p$. The $q(\cdot)$ function captures the idea that increased weapons stocks in either country make any conflict potentially more dangerous. The $f(\cdot)$ function captures the idea that a country's likelihood of winning a conflict is higher, the more weapons it possesses and the fewer weapons possessed by its adversary. Accumulation that increases one country's probability of winning a conflict decreases the probability that the other wins, and each country's probability of winning equals $\frac{1}{2}$ when the countries possess equal weapons stocks. We assume that $f(\cdot)$ and $q(\cdot)$ are continuous and twice-differentiable. Provided that annihilation does not occur, the agents in the victorious country receive utility when they are old equal to $V(\cdot)(1+\alpha)$, and the agents on the losing side have utility when they are old equal to $V(\cdot)(1-\alpha)$, where α is a positive parameter.[6]

[1]Note that while young agents would like to levy lump-sum taxes on the old, such a policy would not be achievable as a Lindahl equilibrium. We are grateful to a referee for clarifying the role of the government in this model.

[2]The term "conflict" should be interpreted liberally; we think of it as any situation in which the two countries engage in contentious behavior, ranging perhaps from a verbal dispute at the United Nations or an expulsion of diplomats, to a confrontational stand-off like the Cuban missile crisis or the Berlin blockade, or even a limited war like Vietnam or Afghanistan.

[3]We assume that this probability is independent of which side initiates; that is, there is no first-mover advantage. In the problem as set out, the sides will never choose to initiate simultaneously.

[4]These last two assumptions together suggest that if a conflict escalates, weapons stocks and all else are annihilated, while in a limited conflict the weapons destroyed are those that would otherwise have depreciated, so the stock bequeathed to the next generation is not affected.

[5]Assume that if p as defined exceeds 1, it takes the value 1, and if p as defined is less than 0, it takes the value 0. Since we restrict attention to symmetric equilibria, these restrictions are never relevant for our analysis.

[6]The augmentation (diminution) of utility as a result of winning (losing) a conflict is analogous to the payment of tribute by the losing side to the winning side. While Michelle R. Garfinkel (1990) makes the transfer of goods explicit, we model the tribute as a psychological gain (loss).

Now consider country A's decision of whether or not to initiate conflict. If B initiates, A's decision is irrelevant. Therefore we need only consider the case in which B does not initiate. Country A's expected utility when old if it initiates is

$$(1-q)p(1+\alpha)V(\cdot) + (1-q)(1-p)(1-\alpha)V(\cdot)$$

and it is $V(\cdot)$ if country A does not initiate. Thus, country A will choose to initiate if and only if

$$p > [1+z(q)]/2$$

where $z(q) = q/[\alpha(1-q)]$. It is indifferent between initiating and not initiating if

$$\eta = -f(\cdot) + [1+z(q)]/2.$$

Similarly, country B is indifferent between initiating and not initiating if

$$1-p = [1+z(q)]/2$$

$$\Rightarrow \eta = 1 - f(\cdot) - [1+z(q)]/2.$$

The probability that neither side initiates is then the probability that η lies between these two values, which simply equals $z(q)$. Since $z(q)$ is increasing in q and $q(\cdot)$ is increasing in both its arguments, it follows that greater weapons accumulation by either side makes conflict less likely to occur, but more likely to be catastrophic if it does occur.[7]

Now consider the first-period weapons choice. Country A chooses weapons (w) and saving (s) to maximize

$$U(c^{y}) + \phi(W^{A}, W^{B})V(c^{o})$$

(subject to the budget constraints discussed below), where

$$\phi(W^{A}, W^{B}) = z(q) + [1-z(q)](1-q) \times \{E(p)(1+\alpha) + [1-E(p)](1-\alpha)\}$$

$$= z(q) + (1-\rho q)[1+2\alpha f(\cdot)] \qquad \rho = (1+\alpha)/\alpha.$$

(Country B solves an analogous problem.) The first term $[z(q)]$ is the probability that neither country initiates. The second term is the probability that someone initiates $(1-z)$, multiplied by the probability that the world nevertheless survives $(1-q)$, multiplied by the expected gain from conflict, which in turn depends upon the expected probability of victory, $E(p) = \frac{1}{2} + f(\cdot)$. In a symmetric equilibrium, $f(\cdot) = 0$, so the expected gain from conflict is zero; in this case $\phi(\cdot)$ can be interpreted as the survival probability. Note that, since $f(\cdot)$ and $q(\cdot)$ are continuous and differentiable, so too is $\phi(\cdot)$. The behavior of the economy depends upon the properties of $\phi(\cdot)$, which are discussed in detail in Section III.

C. *The Agent's Optimization Problem*

The representative agent in country A (and symmetrically for the representative agent in country B) takes as given the return on storage, R, the stocks of weapons at the beginning of period t, $W^{A}(t-1)$ and $W^{B}(t-1)$, and the weapons choice of country B, $w^{B}(t)$. The agent chooses $c^{Ay}(t)$, $c^{Ao}(t+1)$, $w^{A}(t)$, and $s^{A}(t)$ to maximize

$$(3)\quad U(c^{Ay}(t)) + \phi(W^{A}(t), W^{B}(t))V(c^{Ao}(t+1))$$

subject to

$$(4)\quad c^{Ay}(t) + w^{A}(t) + s^{A}(t) \le e$$

$$(5)\quad c^{Ao}(t+1) \le s^{A}(t)R$$

$$(6)\quad W^{A}(t) = (1-\delta)W^{A}(t-1) + w^{A}(t)$$

and $c^{Ay}(t), c^{Ao}(t+1), w^{A}(t), s^{A}(t) \ge 0$.

[7] For an interior solution to the problem set out here, $z(q)$ must be less than 1 [$\Rightarrow q < \alpha/(1+\alpha)$]. For the present, we assume that this condition holds and discuss its implications later in the paper.

Utility maximization yields the following first-order conditions for an interior optimum:

$$(7) \qquad -U'(\cdot)+\phi(\cdot)V'(\cdot)R=0$$

$$(8) \qquad -U'(\cdot)+\phi_1(\cdot)V(\cdot)=0.$$

The restrictions on the utility function ensure that (7) is always satisfied. A necessary condition for (8) is that ϕ_1 be positive; that is, an increase in w^A benefits country A. The assumptions of the model so far are not sufficient to ensure that ϕ_1 is always positive. If ϕ_1 is negative, agents will not invest in weapons, and as shown in Section III, an equilibrium with zero investment in weapons is possible. Henceforth, we assume that ϕ_1 is positive unless otherwise noted.[8]

The individual equates the marginal utility of consumption in youth with the expected marginal utility of consumption in old age [equation (7)]; ϕR can be thought of as an implicit interest rate. Equations (7) and (8) together reveal that the individual also equates the return on saving to the return on weapons accumulation:

$$\phi_1 V=\phi R V'.$$

The second-order conditions from individual maximization require that

$$\tilde{\Delta}\equiv \det\begin{bmatrix} U''+\phi V''R^2 & U''+\phi_1 V'R \\ U''+\phi_1 V'R & U''+\phi_{11}V \end{bmatrix}>0$$

$$\Rightarrow U''\left(\phi V''R^2-\phi_1 V'R\right)$$

$$+\phi_{11}V\left(U''+\phi V''R^2\right)$$

$$-\phi_1 V'RU'\left[\frac{U''}{U'}+\frac{\phi_1}{\phi}\right]>0.$$

Sufficient for this is that $U''/U'+\phi_1/\phi<0$ and $\phi_{11}<0$.

[8]We assume throughout that destruction of weapons in excess of natural depreciation is not possible.

II. The Steady State

A symmetric steady-state Nash equilibrium is given by the quintuple for country A, $\{c^y, c^o, w, W, s\}$, and the identical quintuple for country B, $\{c^y, c^o, w, W, s\}$, such that the following statements hold.

(i) Agents maximize (3) subject to (4)–(6), given R and W.

(ii) The goods market clears:

$$(9) \qquad [c^y+w+s]+c^o=e+sR.$$

(iii) The stock of weapons is constant:

$$(10) \qquad \delta W=w.$$

The equilibrium can be characterized by the two first-order conditions, (7) and (8), and constant weapons stocks, (10). Total differentiation of the first-order conditions evaluated at an interior steady state yields the following system:

$$(11) \quad \begin{bmatrix} u+v & u-a \\ u-\zeta & u-b \end{bmatrix}\begin{bmatrix} ds \\ dw \end{bmatrix}$$

$$=\begin{bmatrix} (1-sv)/R & u & -aw/\delta & \phi_\alpha/\phi \\ \zeta s/R & u & -bw/\delta & \phi_{1\alpha}/\phi_1 \end{bmatrix}\begin{bmatrix} dR \\ de \\ d\delta \\ d\alpha \end{bmatrix}$$

where

$$u\equiv -U''/U'>0$$

$$v\equiv -V''R/V>0$$

$$\zeta\equiv \phi_1/\phi>0$$

$$a\equiv(\phi_1+\phi_2)/\delta\phi\gtrless 0$$

$$b\equiv(\phi_{11}+\phi_{12})/\delta\phi_1\gtrless 0$$

$$\phi_\alpha\equiv \partial\phi/\partial\alpha$$

$$\phi_{1\alpha}\equiv \partial\phi_1/\partial\alpha.$$

For stability the determinant of the left-

hand-side matrix, Δ, must be positive. This implies that

$$\Delta \equiv u(v+\zeta)+a(u-\zeta)-b(u+v)>0.$$

A sufficient condition for $\Delta>0$ is $a(u-\zeta)>0$ and $b<0$.

III. Equilibria

A. *Properties of* $\phi(\cdot)$

We begin this section by discussing the properties of $\phi(\cdot)$, since these are central to a characterization of the model's equilibria. Recall that $\phi(\cdot)$ is the weight attached to old agents' utility and depends upon both the probability of survival to old age and the gain or loss from any conflict that might occur:

$$\phi(W^{A},W^{B})=z(q)+(1-\rho q)(1+2\alpha f).$$

Recall also that $z(q)=q/[\alpha(1-q)]$ so that $z'>0$, $z''>0$. We restrict our attention to symmetric equilibria, in which $q_1=q_2$, $f_1=-f_2$, and $f=0$. Then,

$$(12)\quad \phi_1=(z'-\rho)q_1+2\alpha(1-\rho q)f_1 -2\alpha\rho f q_1 =(z'-\rho)q_1+2\alpha(1-\rho q)f_1$$

$$(13)\quad \phi_2=(z'-\rho)q_2+2\alpha(1-\rho q)f_2 -2\alpha\rho f q_2 =(z'-\rho)q_2+2\alpha(1-\rho q)f_2$$

$$(14)\quad \frac{d\phi(\cdot)}{dw}=\phi_1+\phi_2=2(z'-\rho)q_1.$$

The first term in equation (12) captures two effects. A higher weapons stock in country A increases the probability of catastrophe if either side initiates ($-\rho q_1$) but also therefore increases the risk associated with initiating, making initiation less likely to occur ($z'q_1$). The second term in equation (12) captures the fact that a higher weapons stock benefits country A in the event of conflict; thus, it is positive. Considering equation (13), note that the first term is identical to the first term of equation (12) and reflects the same intuition. A higher weapons stock in country B is costly to country A, however; hence the second term in equation (13) is negative (since $f_2<0$).

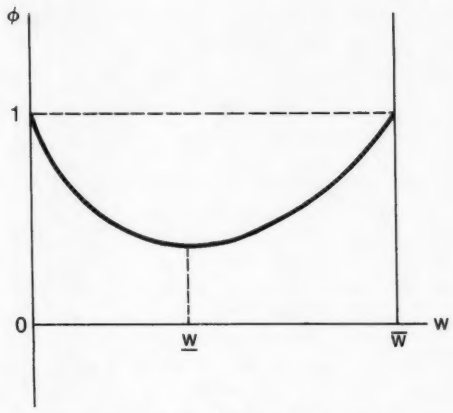


FIGURE 1. THE PROBABILITY OF SURVIVAL FOR SYMMETRIC WEAPONS STOCKS

The behavior of $\phi(\cdot)$, evaluated at a steady state (with $w=\delta W$), is illustrated in Figure 1 under the assumption of symmetry across countries: $w^{A}=w^{B}=w$. At $w=0$, $\phi(0,0)=1$ because $q(0,0)=0$, $z(0)=0$, and $f(0,0)=0$. That is, when neither country has any weapons, conflict is certain but costless because annihilation is impossible. Thus, the probability of survival is unity. As w increases, the danger inherent in any conflict increases, but the likelihood of conflict falls. The first effect dominates for low weapons stocks; the second effect dominates at high weapons stocks. The probability of survival is also unity if agents never initiate conflict because the associated risk of annihilation is so great. Specifically, let $\overline{w}$ be defined by $q(\overline{w}/\delta,\overline{w}/\delta)=1/\rho$. Then $z(q)=1$ and $\phi(\overline{w}/\delta,\overline{w}/\delta)=1$. This possibility resembles the doctrine of "mutually assured destruction" under which deterrence works perfectly because initiation of a conflict carries too great a risk of annihilation.

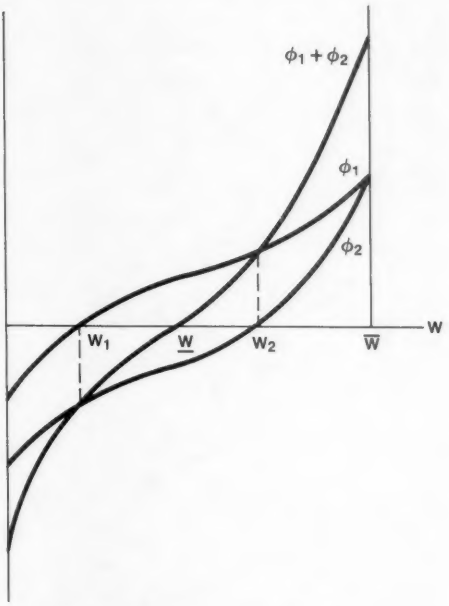


FIGURE 2. DERIVATIVES OF THE SURVIVAL PROBABILITY FUNCTION

Figure 2 shows the partial derivatives of $\phi(\cdot)$ [for arbitrary assumptions on the second derivatives of $\phi(\cdot)$]. It is easily confirmed that at $w=0$, $z'(\cdot)-\rho=-1$, implying that $\phi_1(\cdot)\lesseqgtr 0$, $\phi_2(\cdot)<0$, and $\phi_1(\cdot)+\phi_2(\cdot)<0$. Figure 2 is drawn such that $\phi_1(\cdot)<0$ at $w=0$. At $w=\bar{w}$, $z'(\cdot)-\rho=1+\alpha>0$, implying that $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are positive and equal. From our assumptions on $f(\cdot)$ and $q(\cdot)$, $\phi_1(\cdot)$ lies strictly above $\phi_2(\cdot)$ for $w<\bar{w}$. We thus divide the range of w into four regions. In the region $0<w<w_1$, both ϕ_1 and ϕ_2 are negative. (Since ϕ_1 may be positive for all w, this region may be empty.) For $w_1<w<\underline{w}$, ϕ_1 is positive, ϕ_2 is negative, and their sum $(\phi_1+\phi_2)$ is negative. The function $\phi(\cdot)$ attains its minimum at $\underline{w}$, where $\phi_1+\phi_2=0$. In the region $\underline{w}<w<w_2$, ϕ_1 is positive, ϕ_2 is negative, and their sum is positive. Finally, for $w_2<w<\bar{w}$, both ϕ_1 and ϕ_2 are positive.

Figures 1 and 2 are drawn such that $\phi(\cdot)$ is single-troughed and $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are monotonically increasing in w; this is not necessarily implied by our assumptions on $f(\cdot)$ and $q(\cdot)$, as can be seen by examining equations (15)–(18):

$$(15)\quad \phi_{11}=(z'-\rho)q_{11}+2\alpha(1-\rho q)f_{11}+z''(q_1)^2-4\alpha\rho f_1q_1$$

$$(16)\quad \phi_{12}=(z'-\rho)q_{12}+2\alpha(1-\rho q)f_{12}+z''q_1q_2-2\alpha\rho f_2q_1-2\alpha\rho f_1q_2=(z'-\rho)q_{12}+2\alpha(1-\rho q)f_{12}+z''q_1q_2$$

$$(17)\quad \phi_{11}+\phi_{12}=2\left[(z'-\rho)q_{11}+\alpha(1-\rho q)\times(f_{11}+f_{12})+z''(q_1)^2-2\alpha\rho f_1q_1\right]$$

$$(18)\quad \phi_{22}=(z'-\rho)q_{22}+2\alpha(1-\rho q)f_{22}+z''(q_2)^2-4\alpha\rho f_2q_2.$$

From equations (15)–(18), ϕ_{11}, ϕ_{22}, and ϕ_{12} can all be of either sign, as can the sum $(\phi_{11}+\phi_{12})$.

B. *Steady-State Equilibria*

Our primary concern is with the equilibrium weapons stock and associated survival probability. We thus let $s(w)$ be implicitly defined by the steady-state version of the first-order condition (7):

$$-U'(e-s-w)+\phi(w/\delta, w/\delta)V'(sR)R=0.$$

Increases in w increase the marginal utility of first-period consumption; this effect tends to increase saving. Below $\underline{w}$, increases in w decrease $\phi(\cdot)$, which tends to discourage saving; the opposite is true above $\underline{w}$.[9] Hence $s'(w)$ is negative below $\underline{w}$ and is of ambiguous sign above $\underline{w}$.[10]

[9]The effects of changes in the survival probability on saving in our model are the theoretical counterparts of Joel B. Slemrod's (1986, 1990) finding that an increased threat of nuclear war is associated with decreased saving. We consider the relationship between our results and Slemrod's further in Section IV.

[10]Specifically, $s'(w)=(a-u)/(u+v)\gtreqless 0$.

The two first-order conditions can be combined to yield

$$(19)\quad \phi_1(\cdot)/\phi(\cdot) = V'(\cdot)R/V(\cdot).$$

Steady-state equilibria for this economy can be represented by graphing the two sides of equation (19) as functions of the steady-state weapons stock, w. The right-hand side of (19) is everywhere positive, and it slopes upward if and only if $s'(w) < 0$.

Figures 3 and 4 illustrate the cases of a single interior equilibrium and multiple interior equilibria, respectively. For purposes of exposition, we make the simplifying assumptions that $\phi_1(\cdot)/\phi$ and $\phi_2(\cdot)/\phi$ are increasing in w, and for diagramatic simplicity, we draw them as linear. While a less well-behaved function might divide the range of w into more than four regions, each region would still have the characteristics of one of the four described above. The equilibrium in Figure 3 is stable, as are the two outer equilibria in Figure 4.[11]

Since the right-hand side of (19) is everywhere positive, an interior equilibrium is not possible in the region $0 < w < w_1$. A corner solution with zero weapons accumulation is possible, however, if $V'R/V$ lies above ϕ_1/ϕ at $w = 0$ (Fig. 5), so that the nonnegativity constraint on weapons accumulation is binding. Note that if $\phi_1(0,0) < 0$, then this corner solution must be an equilibrium.

[11]Two distinct types of multiplicity may arise. First, there may be multiple steady states, but the game between countries may still have a unique equilibrium at any given date. Such multiplicity can be attributed to strategic complementarities between past generations' accumulation decisions and the present generations' decisions ($\phi_{11} + \phi_{12} > 0$): the more weapons that were accumulated in the past, the greater the return to accumulation in the present. The final steady state then depends upon initial weapons stocks. Alternatively, there may be multiple equilibria in a given time period, with the associated possibility of coordination failure. For example, each country might devote a lot of its endowment to weapons simply because the other is doing so, even if there is a better Nash equilibrium with less investment in weapons. This possibility arises if there are strategic complementarities within the period ($\phi_{12} > 0$): the more weapons accumulated by one country, the higher the return to weapons accumulation in the other country. See Russell W. Cooper and John (1988) for more discussion of strategic complementarities and coordination failures.

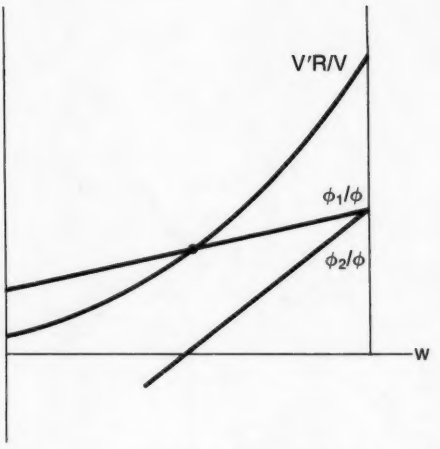


FIGURE 3. A UNIQUE STEADY-STATE EQUILIBRIUM

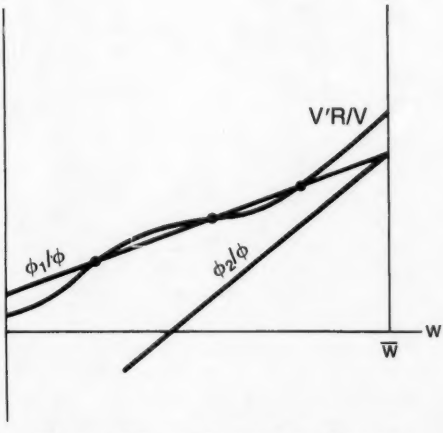


FIGURE 4. MULTIPLE STEADY-STATE EQUILIBRIA

Another possible corner solution has agents accumulating weapons to such an extent that initiating conflict is never beneficial and so will never occur because the risk of annihilation is too great; that is, weapons are accumulated up to $\overline{w}$. In this no-conflict equilibrium, ϕ_1/ϕ lies above $V'R/V$ at $w = \overline{w}$, as is shown in Figure 6.

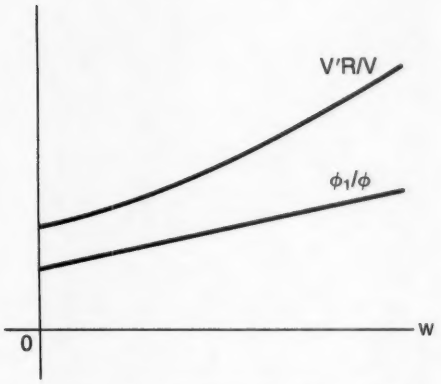


FIGURE 5. AN EQUILIBRIUM WITHOUT WEAPONS ACCUMULATION

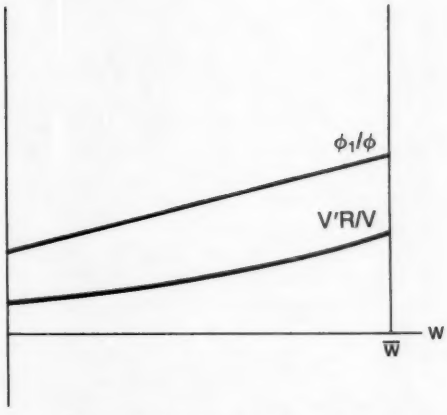


FIGURE 6. A NO-CONFLICT EQUILIBRIUM

Arms races, such as have been observed throughout history, can be interpreted in our model as a dynamic adjustment to a steady state with positive accumulation. Starting from some weapons stock below equilibrium, successive generations in each country engage in weapons investment that augments the weapons stock. Given one country's investment in weapons, the self-interest of the other country dictates that it should also engage in investment. Such investment is socially counterproductive below $\underline{w}$, in the sense that symmetric changes in weapons stocks simply decrease the survival probability without providing an advantage to either country. The consequence is that, below $\underline{w}$, each generation bequeaths a more dangerous world to generations that follow. If the steady-state weapons stock exceeds $\underline{w}$, then the post-$\underline{w}$ buildup actually makes the world safer because of the deterrent effect of possible annihilation.

IV. Comparative Statics

Now consider how steady-state saving and weapons accumulation are affected by changes in the exogenous variables of the model. In the following comparative-static experiments, we restrict our attention to symmetric interior stable equilibria; thus we consider the symmetric response of saving and weapons accumulation to, for example, an equal change in the return on storage in both countries. That is, we analyze the effects on equilibrium of changes in the characteristics of the world.[12] Many of the results are ambiguous, reflecting the facts that the effective interest rate, $\phi(\cdot)R$, is endogenous even though the return on saving is fixed and that the assumptions of our model place few restrictions on the second-derivative properties of $\phi(\cdot)$.

PROPOSITION 1: *(i) Increases in the return to storage have an ambiguous effect on saving and weapons accumulation. (ii) If $V(\cdot)$ is logarithmic, then increases in the return to storage increase weapons accumulation but still have an ambiguous effect on saving.*

PROOF:

(i) From (11),

$$\frac{ds}{dR} = \frac{1}{\Delta R}[(b-u)(sv-1)+s\zeta(a-u)] \gtrless 0$$

$$\frac{dw}{dR} = \frac{1}{\Delta R}[u(sv-1)+\zeta(su+1)] \gtrless 0.$$

[12]The comparative-static results for the zero-weapons and zero-conflict equilibria are straightforward, since the weapons stock does not respond to small changes in the parameters.

(ii) If $V(\cdot)$ is logarithmic, then $sv = 1$ and

$$\frac{ds}{dR} = \frac{1}{\Delta R}[s\zeta(a-u)] \lesseqgtr 0$$

$$\frac{dw}{dR} = \frac{1}{\Delta R}[\zeta(su+1)] > 0.$$

When the return on saving rises, consumption in old age is more attractive. Agents can effectively shift consumption into the future in two ways: by increasing future consumption directly through saving, or indirectly by accumulating weapons. An increase in the return on saving increases the value of consumption in old age, encouraging agents to accumulate more weapons to benefit from this higher consumption. For $w < \underline{w}$ $(a < 0)$, this extra accumulation reduces $\phi(\cdot)$ and so reduces the effective return on saving, thus discouraging saving. For $w > \underline{w}$, the saving response is ambiguous. The ambiguity stems partly from the familiar income and substitution effects. In the special case in which $V(\cdot)$ is logarithmic, these effects cancel.

PROPOSITION 2: *Increases in the depreciation rate of the weapons stocks have an ambiguous effect on saving and weapons accumulation.*

PROOF:
From (11),

$$\frac{ds}{d\delta} = \frac{w}{\delta\Delta}[u(b-a)] \gtreqless 0$$

$$\frac{dw}{d\delta} = \frac{w}{\delta\Delta}[a(u-\zeta) - b(v+u)] \gtreqless 0.$$

Increases in δ decrease the aggregate weapons stock (i.e., $dW/d\delta < 0$), affecting both ϕ and ϕ_1, and thus a and b, respectively. For $w > \underline{w}$ the probability of survival, $\phi(w/\delta, w/\delta)$, is increasing in w so that $a > 0$. Additional restrictions are needed to sign $ds/d\delta$ and $dw/d\delta$.

PROPOSITION 3: *Increases in the endowment of the young have an ambiguous effect on saving and increase weapons accumulation.*

PROOF:
From (11),

$$\frac{ds}{de} = \frac{1}{\Delta}[u(a-b)] \gtreqless 0$$

$$\frac{dw}{de} = \frac{1}{\Delta}[u(v+\zeta)] > 0.$$

Increases in the endowment encourage weapons accumulation. The effect on saving depends upon how the increased accumulation of weapons affects the implicit interest rate $\phi(\cdot)R$. If the economy is in the region where weapons accumulation increases world safety $(w > \underline{w})$, then $a > 0$, so the increased accumulation raises the effective return on saving and encourages saving. For $b > 0$, however, increased accumulation increases the marginal return on accumulation (ϕ_1), making saving a relatively less attractive way to increase expected old-age utility.

PROPOSITION 4: *Increases in the return to winning a conflict, α, have an ambiguous effect on both saving and weapons accumulation.*

PROOF:
From (11),

$$\frac{ds}{d\alpha} = \frac{1}{\Delta}\left[\frac{\phi_\alpha}{\phi}(u-b) - \frac{\phi_{1\alpha}}{\phi_1}(u-a)\right] \gtreqless 0$$

$$\frac{dw}{d\alpha} = \frac{1}{\Delta}\left[\frac{\phi_{1\alpha}}{\phi_1}(u+v) - \frac{\phi_\alpha}{\phi}(u-\zeta)\right] \gtreqless 0$$

since $\phi_\alpha < 0$ and $\phi_{1\alpha} \gtreqless 0$.

This result resembles that in Proposition 2, reflecting the fact that changes in δ and changes in α affect agents' decisions through their effect on $\phi(\cdot)$. The resemblence is apparent when it is noted that ϕ_α and $\phi_{1\alpha}$

play the same role as a and b in Proposition 2.

An increase in α increases the likelihood of initiation and hence reduces the survival probability (i.e., $\phi_\alpha < 0$). Such a change can be interpreted as a decline in inherent world safety. The increased likelihood of annihilation has the direct effect of reducing saving, consistent with Slemrod's (1986, 1990) findings that the increased threat of nuclear war is accompanied by a decrease in saving. Other factors in our model can make the overall impact of a change in α be of either sign, however. In particular, changes in α have an ambiguous effect on the marginal return to accumulation ($\phi_{1\alpha} \gtrless 0$) and so may encourage or discourage saving through this channel.

William Shepherd (1988) argued that the United States and the (former) Soviet Union were inherently safe countries, in the sense that neither had much to gain from initiating conflict. He therefore concluded that weapons accumulation in these countries was inconsistent with the needs of deterrence. In our model, however, the ambiguous effect of α on ϕ_1 permits high weapons accumulation to arise in inherently safe economies (when $\phi_{1\alpha} < 0$). Nevertheless, if decreases in safety increase the return to accumulation, so that $\phi_{1\alpha} > 0$, and if $u - \zeta > 0$, then we find that weapons accumulation is lower in safer worlds.

V. Welfare Analysis

One reason for adopting an overlapping-generations framework is that it permits explicit analysis of intergenerational welfare effects. In this model there are three distinct sources of externality. First, contemporaneous weapons accumulation in country A affects the welfare of agents in country B, by its effect both on the likelihood of conflict and the expected gain from any conflict that does occur. Second, since weapons outlast the agents who invest in them, the accumulation decisions of previous generations in country A affect the weapons stock and, hence, the utility of the current generation in country A. Third, the accumulation decisions of previous generations of country A affect the utility of the current generation in country B. These externalities differ depending upon the equilibrium weapons stock. The presence of externalities leads to the conclusion that the Nash equilibrium may not be Pareto-efficient.

The only link between generations in this model is through the effect of weapons accumulation, so the depreciation rate of the weapons stock, δ, provides a measure of the extent of the intergenerational externality. In particular, notice that if $\delta = 1$, the link between generations disappears, and there is no intergenerational externality; successive generations simply play a series of unconnected one-shot games. The appropriate magnitude of δ depends on how quickly weapons become obsolete and on the length of a period.

A. *First Best*

The first-best symmetric outcome in this model arises if a social planner can prevent the initiation of conflict. At any positive symmetric weapons stocks, conflict is a negative-sum game by the assumptions on $f(\cdot)$ and $q(\cdot)$. If the planner can enforce noninitiation, moreover, there is no reason to accumulate weapons, so a zero stock of weapons is first-best. We henceforth assume that noninitiation is never enforceable and so restrict attention to the effects of resource reallocations. Not surprisingly, even in the absence of enforceable noninitiation, the first-best symmetric allocation in this model entails zero accumulation of weapons, since this implies a survival probability of unity. Complete multilateral disarmament starting at some date τ when agents possess positive weapons stocks [i.e., setting $w(\tau) = 0$ in both countries for all $t \geq \tau$] need not be Pareto-improving, however. These results are stated as Proposition 5.

PROPOSITION 5: *(i) Zero weapons accumulation ($w = 0$) is a necessary condition for a first-best symmetric equilibrium. (ii) If $\underline{w} \geq w(\tau) > 0$, or if δ is sufficiently large, then complete disarmament [$w(t) = 0 \ \forall \ t \geq \tau$] is Pareto-improving; if $\underline{w} < w(\tau)$ and δ is small, complete disarmament will be Pareto-*

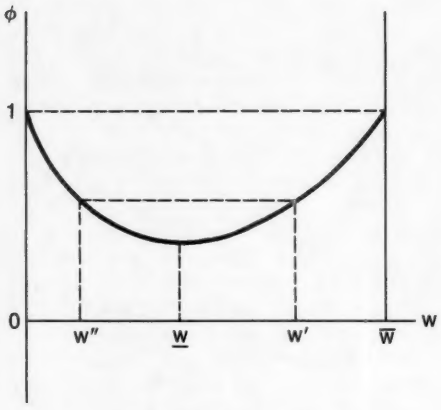


FIGURE 7. THE FIRST-BEST EQUILIBRIUM

improving if and only if

$$U(e-\underline{w}-s(\underline{w}))+\phi(\underline{w}/\delta,\underline{w}/\delta)V(s(\underline{w})R)$$

$$\geq U(e-w(\tau)-s(w(\tau)))$$

$$+\phi(w(\tau)/\delta,w(\tau)/\delta)V(s(w')R)$$

where $s(\cdot)$ *indicates the optimal choice of saving for a given* w.

PROOF:

(i) Consider first any $w' > \underline{w}$ as shown in Figure 7. Then by the continuity of $\phi(\cdot)$ there exists a $w'' < \underline{w}$ such that $\phi(w''/\delta, w''/\delta) = \phi(w'/\delta, w'/\delta)$. At w'', the probability of survival is the same as at $w = w'$, but fewer resources are devoted to weapons, so consumption is higher, and all agents can be made better off. Now consider w such that $\underline{w} \geq w > 0$. Then a decrease in w increases the survival probability and increases resources available for consumption. By the continuity of $\phi(\cdot)$ and $\phi(0,0) = 1$, the first-best equilibrium occurs at $w = 0$.

(ii) If the initial weapons stock is below $\underline{w}$, then weapons depletion increases welfare, as noted in (i). If the initial weapons stock exceeds $\underline{w}$ and δ is small, disarmament initially decreases world safety by encouraging initiation of conflict and so will only be Pareto-improving if the resources freed by nonaccumulation outweigh the cost of a lower survival probability. In particular, if $\delta \approx 0$, this must be true at $\underline{w}$, which is the condition stated in the proposition. If δ is large, however, the economy may "jump over" the minimum point $\underline{w}$; indeed, if $\delta = 1$, multilateral disarmament immediately moves the economy to its first-best equilibrium.

The first-best equilibrium in this model is identical to that in van der Ploeg and de Zeeuw's (1990) model. The overlapping-generations framework, however, complicates the welfare analysis of disarmament because different generations are affected differently. Proposition 5 illustrates one important conclusion of this paper: if prior accumulation decisions have placed the economy above $\underline{w}$, then intergenerational externalities complicate any attempts at Pareto-improving disarmament (presuming, of course, that weapons cannot be costlessly destroyed).

As a straightforward corollary of Proposition 5, we note that if the economy exhibits multiple symmetric equilibria in the range $[0,\underline{w}]$, then these equilibria are inversely Pareto-ranked by the stock of weapons. Further, if the corner solution $w = 0$ is an equilibrium, then it is evidently Pareto-superior to any other equilibrium. Henceforth, we abstract from the possibility of multiple equilibria.

B. *Pareto Improvements with Intergenerational Transfers*

Proposition 5 shows that disarmament need not be Pareto-improving, since it may be costly to early generations. Pareto improvements are generically possible if we permit intergenerational transfers, but such improvements may involve additional accumulation of weapons rather than disarmament. First, note that we can write steady-state utility as a function of steady-state w

as follows:

$$\Upsilon(w) = U(e - w - s(w)) + \phi(w/\delta, w/\delta)V(s(w)R)$$

where $s(w)$ indicates optimal saving. We know from Proposition 5 that this has a global maximum at $w = 0$. Differentiating and appealing to the envelope theorem, we obtain

$$\Upsilon'(w) = -U'(\cdot) + [(\phi_1 + \phi_2)/\delta]V(\cdot) = (V(\cdot)/\phi) \times [(\phi_1 + \phi_2)/\phi\delta - V'(\cdot)R/V(\cdot)]$$

using the first-order condition $U'(\cdot) = \phi R V'(\cdot)$.

If the economy is initially in noncooperative equilibrium, then $\phi_1/\phi = V'(\cdot)R/V$, implying

$$\Upsilon'(w) = V(\cdot)[(\phi_1/\phi)(1-\delta)/\delta + (\phi_2/\phi)/\delta] = \frac{V(\cdot)}{\phi}\left[(\phi_1 + \phi_2)\left(\frac{1-\delta}{\delta}\right) + \phi_2\right].$$

The first term is the intergenerational externality arising from higher inherited weapons stocks from both countries, while the second term is the within-generation across-country externality. The intergenerational externality is negative (i.e., higher weapons stocks reduce the welfare of future generations) for $w < \underline{w}$, and it is positive for $w > \underline{w}$. The within-generation externality is negative for $w < w_2$ and positive for $w > w_2$. Since $w_2 > \underline{w}$, it follows that both externalities are negative if the economy is in a noncooperative equilibrium below $\underline{w}$. Higher steady-state weapons stocks in this region are thus associated with lower welfare. Conversely, both externalities are positive if the equilibrium entails a weapons stock in excess of w_2, in which case higher equilibrium weapons stocks are associated with higher welfare. Propositions 7 and 8 in Section VI show that, in both these cases, Pareto improvements are possible without intergenerational transfers.

In the intermediate region, the intragenerational externality is negative, but the intergenerational externality is positive. If $\Upsilon'(w) > 0$ at a noncooperative equilibrium, then the intergenerational externality dominates, and higher steady-state weapons stocks are associated with higher utility, suggesting that buildup of weapons would be Pareto-improving. Such a conclusion is not quite right, however, because weapons buildup imposes costs on current generations but provides benefits to future generations. As established in Proposition 6, Pareto-improving buildup is possible if

$$\phi_1(1-\delta) + \phi_2 R = (\phi_1 + \phi_2)[(1-\delta)/\delta] + \phi_2[1 + (R-1)/\delta] > 0.$$

This expression differs from $\Upsilon'(w)$ by the addition of $\phi_2(R-1)/\delta < 0$. A high return to storage or a low depreciation rate thus means disarmament may be Pareto-improving even if $\Upsilon'(w) > 0$.

PROPOSITION 6: *Assume that the economy is dynamically efficient ($R > 1$) and in a noncooperative equilibrium (w^{NC}) with $\phi_1 + \phi_2 > 0$ and $\phi_2 < 0$ (i.e., $\underline{w} < w^{NC} < w_2$).*[13] *Pareto improvements are generically possible.*

PROOF:

We first show that if $\phi_1(1-\delta) + \phi_2 R > 0$, then a small increase in weapons is Pareto-improving with appropriate transfers. Let $w(t) = w^{NC}\ \forall\ t \neq \tau$; $w(\tau) = w^{NC} + \varepsilon, \varepsilon$ small. Then, starting at date $\tau + 1$, construct transfers $T(t)$ from the young of generation t to the old of generation $t-1$ (i.e., a transfer from young to old at time t) such that the expected utility of generation $t-1$ is unchanged. Recall that

$$\phi(W(t+1), W(t+1)) = \phi(W(t)(1-\delta) + w(t), W(t)(1-\delta) + w(t)).$$

[13]These three assumptions rule out the obvious sources of Pareto improvements. See Propositions 7 and 8 for more discussion.

Take a Taylor-series expansion around $w = w^{NC}$ and $T(\tau+1)=0$ of the expected utility of generation τ. Using the first-order conditions, the change in expected utility is then

$$\phi_2(\cdot)V(\cdot)\varepsilon + \phi(\cdot)V'(\cdot)T(\tau+1).$$

Setting $T(\tau+1)$ so that expected utility is unchanged implies

$$T(\tau+1) = -\phi_2(\cdot)V(\cdot)\varepsilon/\phi(\cdot)V'(\cdot) > 0.$$

Generation τ is made worse off by the extra weapons accumulation of the other country ($\phi_2 < 0$) and so must be compensated when old by the next generation of young agents.

Similarly, $T(\tau+2)$ can be defined so that the expected utility of generation $\tau+1$ is unchanged:

$$\begin{aligned}
&(\phi_1+\phi_2)(1-\delta)V(\cdot)\varepsilon - U'(\cdot)T(\tau+1) \\
&\quad + \phi V'(\cdot)T(\tau+2) = 0 \\
&\Rightarrow (\phi_1+\phi_2)(1-\delta)V(\cdot)\varepsilon \\
&\quad - \phi R V'(\cdot)T(\tau+1) \\
&\quad + \phi V'(\cdot)T(\tau+2) = 0 \\
&\Rightarrow [(\phi_1+\phi_2)(1-\delta)+\phi_2 R]V(\cdot)\varepsilon \\
&\quad + \phi V'(\cdot)T(\tau+2) = 0.
\end{aligned}$$

Likewise, $T(\tau+3)$ is defined by

$$\begin{aligned}
&(\phi_1+\phi_2)(1-\delta)^2 V(\cdot)\varepsilon - U'(\cdot)T(\tau+2) \\
&\quad + \phi V'(\cdot)T(\tau+3) = 0 \\
&\Rightarrow (\phi_1+\phi_2)(1-\delta)^2 V(\cdot)\varepsilon \\
&\quad - \phi R V'(\cdot)T(\tau+2) \\
&\quad + \phi V'(\cdot)T(\tau+3) = 0 \\
&\Rightarrow \{(\phi_1+\phi_2)[(1-\delta)^2+(1-\delta)R]+\phi_2 R^2\}V(\cdot)\varepsilon \\
&\quad + \phi V'(\cdot)T(\tau+3) = 0.
\end{aligned}$$

By repeated substitution, $T(\tau+N)$ is defined by

$$\begin{aligned}
&\left[(\phi_1+\phi_2)\sum_{n=1}^{N}(1-\delta)^n R^{N-n} + \phi_2 R^N\right] \\
&\quad \times V(\cdot)\varepsilon + \phi V'(\cdot)T(\tau+N) = 0 \\
&\Rightarrow \left\{(\phi_1+\phi_2)\sum_{n=1}^{N}[(1-\delta)/R]^n + \phi_2\right\} \\
&\quad \times R^N V(\cdot)\varepsilon + \phi V'(\cdot)T(\tau+N) = 0.
\end{aligned}$$

As $N \to \infty$, and since $(1-\delta)/R < 1$, the term in braces tends to

$$\frac{(\phi_1+\phi_2)(1-\delta)}{R-(1-\delta)} + \phi_2$$

which is positive since $\phi_1(1-\delta)+\phi_2 R > 0$ by assumption. It follows that, at some date τ^*, the transfers change sign; that is, members of generation τ^*-1 make both a positive transfer when young to the old of generation τ^*-2 and a positive transfer when old to the young of generation τ^*.

Now set $T(t)=0\ \forall\ t \geq \tau^*$. The welfare of all generations up to and including generation τ^*-2 is unchanged by construction. Generation τ^*-1 is better off since it does not have to make the positive transfer at τ^* that would leave its utility unchanged. All future generations are also better off because they neither make nor receive transfers but do inherit a larger weapons stock. This pattern of transfers is thus Pareto-improving. A similar argument holds if $\phi_1(1-\delta)+\phi_2 R < 0$; in this case a small decrease in $w(\tau)$ is Pareto-improving.

Proposition 6 establishes that there is some critical value of the weapons stock above which Pareto improvements are (locally) possible by building up weapons and below which Pareto improvements are (locally) possible by reduction of weapons stocks. This critical value is higher, the higher the return on storage and the depreciation rate.

VI. Cooperative Equilibria

A. *Cooperation Across Countries*

Consider the possibility that members of a given generation in countries A and B cooperate. As in the earlier discussion, we disregard the trivial case of a nonaggression pact whereby agents agree not to initiate conflict (note that such an agreement in general will not be time-consistent) and restrict our attention to the cooperative choice of weapons. We think of such cooperation as a verifiable arms treaty that neither restricts future generations nor precludes conflict in the current period. We assume, in keeping with our symmetry assumptions, that both countries have equal bargaining power. The problem faced by the two countries is to choose both their weapons stocks to maximize (3) subject to (4) and (5), taking as given the inherited weapons stocks in both countries. The first-order conditions for an interior equilibrium in this problem are

$$(20) \qquad -U'(\cdot)+\phi(\cdot)V'(\cdot)R=0$$

$$(21) \qquad -U'(\cdot)+(\phi_1+\phi_2)V(\cdot)=0.$$

These first-order conditions are identical to those from the original problem [equations (7) and (8)], except that there is an additional $\phi_2 V(\cdot)$ term in equation (21). This indicates that the two countries internalize the intragenerational externality. The intergenerational externality remains, however. We consider this case to be of particular interest since intragenerational externalities are intrinsically easier to internalize than external effects across generations.

We now consider the possibility of a regime change at time τ. That is, we suppose that the economy is at a noncooperative steady-state equilibrium (w^{NC}) prior to τ, and that at all dates including and after τ, both countries cooperate on their choice of weapons. Such a regime change could be interpreted as corresponding to the end of the Cold War. The following propositions reveal that both buildup and reduction of weapons stocks are possible responses and that disarmament need not be Pareto-improving.[14]

PROPOSITION 7: *If $\phi_2>0$ at w^{NC} so that $\bar{w}>w^{NC}>w_2$, then agents in both countries will choose cooperatively to build up their stocks of weapons, either to an interior equilibrium, w^C, or to $\bar{w}$. Such buildup is Pareto-improving.*

PROOF:

First consider accumulation to an interior cooperative equilibrium, w^C, as illustrated in Figure 8. Let $\tilde{w}(t)=\delta W(t)$. From equation (2),

$$\tilde{w}(t+1)=(1-\delta)\tilde{w}(t)+\delta w(t).$$

If, at time t, $w^{NC}\le\tilde{w}(t)<w^C$, then the following inequality holds:

$$\frac{\phi_1(\tilde{w}(t)/\delta,\tilde{w}(t)/\delta)+\phi_2(\tilde{w}(t)/\delta,\tilde{w}(t)/\delta)}{\phi(\tilde{w}(t)/\delta)} > \frac{V'(s(\tilde{w}(t))R)R}{V(s(\tilde{w}(t))R)}.$$

At date t, agents choose their weapons stock, $w(t)$, to equate $(\phi_1+\phi_2)/\phi$ and $V'R/V$, where $V(\cdot)$ and $\phi(\cdot)$ are functions of $w(t)$ and the inherited weapons stock, $\tilde{w}(t)$, which they take as given. If $\delta<1$, then $(\phi_1+\phi_2)/\phi$ responds less to changes in $w(t)$ than to changes in steady-state w, implying that the out-of-steady-state $(\phi_1+\phi_2)/\phi$ passes through $\tilde{w}(t)$ and is flatter than the steady-state relationship. (In the limit as $\delta\to 0$, this curve becomes horizontal.) Also, if $\delta<1$, changes in $w(t)$ lead to a smaller increase in $\phi(\cdot)$ than do changes in steady-state w and so have a smaller effect on

[14]The analysis of Proposition 7 uses our assumption that $(\phi_1+\phi_2)/\phi$ is upward-sloping. If this relationship slopes downward, but $V'R/V$ (for an individual generation) still slopes upward, then the analysis is almost identical. In this case individual generations might choose weapons investment in excess of w^C, the cooperative equilibrium, but the aggregate weapons stock still will not exceed w^C. If both relationships slope downward, then we cannot rule out the possibility of oscillatory adjustment of the weapons stock.

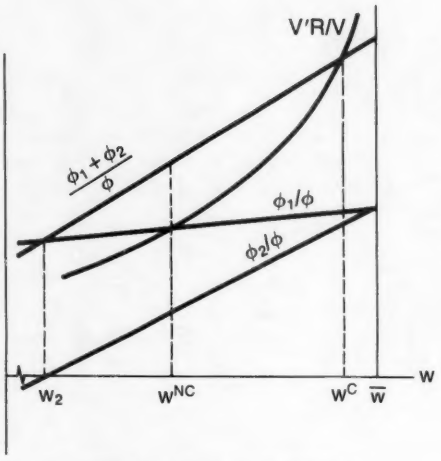


FIGURE 8. BUILDUP TO AN INTERIOR COOPERATIVE EQUILIBRIUM

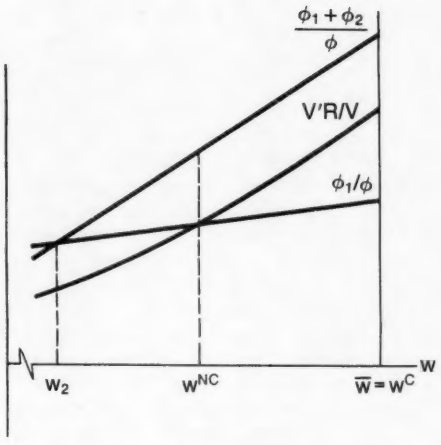


FIGURE 9. BUILDUP TO A NO-CONFLICT EQUILIBRIUM

saving through this channel. As a consequence, the out-of-steady-state $V'R/V$ rotates through $\tilde{w}(t)$ and becomes steeper than its steady-state counterpart. By continuity of $V(\cdot)$ and $\phi(\cdot)$, it follows that

$$\tilde{w}(t) < w(t) < w^C$$

$$\Rightarrow \tilde{w}(t) < \tilde{w}(t+1) < w^C.$$

If $\delta = 1$, then there is no intergenerational externality, and the out-of-steady-state and steady-state relationships are coincident. Thus,

$$w(t) = \tilde{w}(t+1) = w^C.$$

At the time of the regime change $\tilde{w}(\tau) = w^{NC}$. Therefore if $\delta < 1$, the two countries cooperatively build up their weapons stocks, which asymptotically approach w^C. If $\delta = 1$, the two countries move immediately to w^C.

Buildup to $\bar{w}$ is illustrated in Figure 9. The analysis is similar, except that individual generations may choose investment in excess of $\bar{w}$, and so $\bar{w}$ is obtained in finite time even if $\delta < 1$. Buildup is Pareto-improving since $\phi_1 + \phi_2 > 0$ for $w^{NC} < w < w^C$: a larger world weapons stock bestows a positive externality on future generations. Generations after τ are better off since they could always choose the same allocation of resources as at the noncooperative equilibrium and benefit from the larger inherited weapons stock.

If the economy is originally in equilibrium above w_2, then cooperation entails internalization of a positive within-generation externality, so that agents choose to build up their weapons stocks. When the within-generation externality (ϕ_2) is positive, then so too is the across-generation externality $(\phi_1 + \phi_2)[(1-\delta)/\delta]$. The two countries thus cooperatively agree to forgo more consumption to make conflict less likely, and despite the lower consumption, cooperation is Pareto-improving. This result is reminiscent of Garfinkel's (1990) finding that adversarial countries in a stochastic environment might accumulate weapons in a cooperative equilibrium. Her result obtains because, in general, there do not exist punishments sufficiently harsh to make not deviating from the cooperative agreement incentive-compatible in the absence of accumulation. Similarly, cooperation that leads to disarmament need not be Pareto-improving, as established in the following proposition.

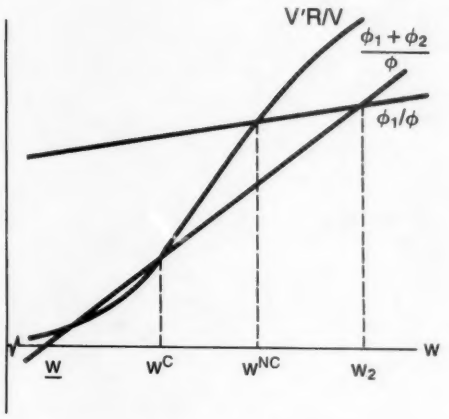


FIGURE 10. REDUCTION OF WEAPONS STOCKS TO AN INTERIOR EQUILIBRIUM

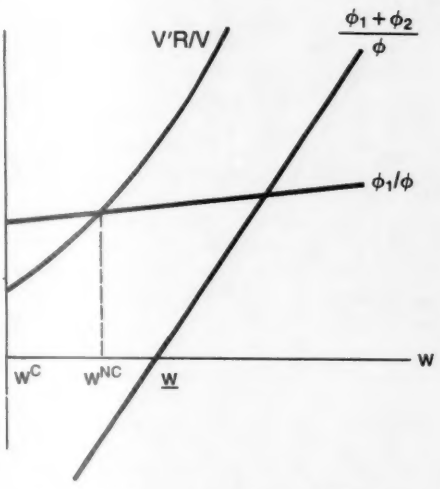


FIGURE 11. REDUCTION OF WEAPONS STOCKS TO AN EQUILIBRIUM WITHOUT WEAPONS ACCUMULATION

PROPOSITION 8: *(i) If* $\phi_2 < 0$ *at* w^{NC} *and* $\underline{w} < w^{NC} < w_2$, *then agents in the two countries will choose cooperatively to reduce weapons stocks to an interior equilibrium or to* $w = 0$. *This reduction need not be Pareto-improving. (ii) If* $\phi_2 < 0$ *at* w^{NC} *and* $w^{NC} < \underline{w}$, *agents in the two countries will cooperatively cease all weapons investment and reduce weapons stocks to* $w = 0$. *This reduction is Pareto-improving.*

PROOF:

(i) Arms reduction to an interior equilibrium is represented in Figure 10; the analysis is analogous to that in Proposition 7. Since $\phi_2 < 0$ at w^{NC}, agents internalize a negative intragenerational externality and so cooperatively choose to reduce their weapons stocks. Since $w^{NC} > \underline{w}$, $\phi_1 + \phi_2 > 0$, and the intergenerational externality is therefore positive. Disarmament in general imposes costs on subsequent generations by decreasing the survival probability and so need not be Pareto-improving. Similar analysis applies for reduction of weapons stocks to zero; in this case, arms reduction is initially costly to future generations but eventually benefits them.

(ii) Reduction of weapons stocks to zero is represented in Figure 11. Since $V'R/V$ is everywhere positive and $(\phi_1 + \phi_2)/\phi$ is negative, generation τ and all subsequent generations immediately move to the corner solution where $w = 0$. Over time, natural depreciation of weapons allows the aggregate weapons stock to decline asymptotically to zero. Below $\underline{w}$, decreases in the aggregate weapons stock increase the survival probability and so impose positive externalities on subsequent generations. Disarmament is therefore Pareto-improving.

Any individual generation does not take account of the effects of its actions on the safety of future generations. Internalization of the within-generation externality thus need not be Pareto-improving for standard second-best reasons: it worsens the across-generation externality. Note, though, that part (i) of Proposition 8 does *not* imply that disarmament is never Pareto-improving even if the noncooperative weapon stock exceeds $\underline{w}$. First, if $\delta = 1$, then there is no intergenerational externality. Second, even if a generation after τ is made worse off because it inherits a lower weapons stock, it also benefits from the fact that it in turn is able to

internalize the intragenerational externality through international cooperation: a cooperative choice of a smaller weapons stock frees resources for consumption.

Our results are consistent with the model of Intriligator and Brito (1984), which demonstrates that a reduction of weapons stocks may promote war rather than peace.[15] The buildup scenario is also consistent with Intriligator and Brito's (1984) finding that buildup may increase world safety, and in our model, it may increase world welfare as well.

B. *Cooperation Across Generations*

Suppose that social planners in each country choose the resource allocation for their individual countries for all time, taking the rules of international interaction as given. This imposes within-country intergenerational cooperation while sustaining the potential for international conflict. In the setting of our model, this is perhaps less plausible in that it requires internalization of externalities between agents who are not alive at the same time. It does however allow for comparison with the infinitely-lived-agent models in the arms-control literature.[16] We analyze this problem as a one-shot game in which a social planner in each country treats all generations symmetrically and chooses a constant allocation of resources through time. In this scenario, the planners can ensure that their countries are dynamically efficient and will impose a larger weapons stock on the world, as shown in Proposition 9.

PROPOSITION 9: *If the intergenerational within-country externality is internalized, then goods will be stored only if storage is dynamically efficient. If the economy is dynamically efficient, the equilibrium weapons stock will be higher than in the noncooperative equilibrium.*

PROOF:

The planner for country A chooses w^{A} and c^{o} to maximize

$$U(c^{y}) + \phi(w^{A}/\delta, w^{B}/\delta)V(c^{o})$$

subject to

$$\begin{aligned} c^{y} &= e - c^{o} - w \quad \text{if } R < 1 \\ c^{o} &= R(e - c^{y} - w) \quad \text{if } R \geq 1. \end{aligned}$$

The first-order conditions yield

$$(22)\quad U'(\cdot) = \begin{cases} \phi V'(\cdot) & \text{if } R < 1 \\ \phi V'(\cdot) R & \text{if } R \geq 1 \end{cases}$$

and

$$(23)\qquad U'(\cdot) = \phi_{1} V(\cdot)/\delta.$$

Equation (22) establishes that storage is only undertaken if it is dynamically efficient, that is, if $R \geq 1$.[17] If $R < 1$, the planner achieves the optimal allocation of goods through transfers from young to old. The only difference between equations (23) and (8) is the $1/\delta$ in (23): ϕ_{1}/δ represents the total effect on country A of present weapons accumulation. As is clear from Figure 12, the equilibrium weapons stock given intergenerational cooperation, w^{IC}, exceeds the noncooperative weapons stock, w^{NC}.[18]

[15] Recall that in our model, lower weapons stocks are *always* associated with a greater likelihood of conflict, but a smaller probability of disaster should conflict occur. This is because the probability of annihilation, $q(\cdot)$, is increasing in w, but the probability that neither side intitiates, $z(q)$, is increasing in $q(\cdot)$. The trade-off between these two effects gives rise to the behavior of the survival probability, $\phi(\cdot)$, as w varies. In comparing our results to those of Intriligator and Brito (1984), we are interpreting "war" in our model as annihilation, not conflict.

[16] Intergenerational cooperation also might be a result of altruism across generations, as in the literature on Ricardian equivalence.

[17] If the planner discounts the welfare of future generations, or if there is population growth, then this condition is amended in the usual way.

[18] If, as can occur, $V'R/V$ is downward-sloping, then it might seem that intergenerational cooperation would reduce weapons accumulation. In this case, however, the equilibria being compared are neither stable nor unique.

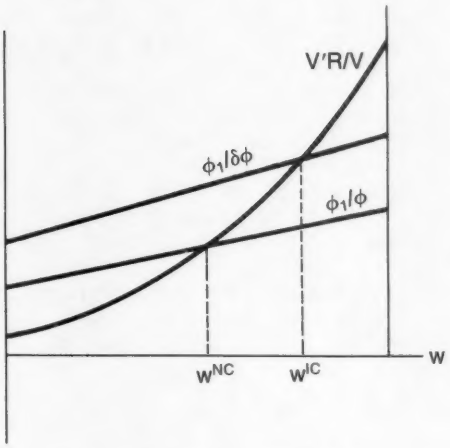


FIGURE 12. AN EQUILIBRIUM WITH INTERGENERATIONAL COOPERATION

Since old agents' weapons accumulation generates uncompensated benefits for young agents in their own country, it is possible that a social contract of the type proposed by Laurence J. Kotlikoff et al. (1988) could be developed to induce each generation to accumulate more weapons. Notice that, while this is welfare-improving from the perspective of a single country, it may actually lower world welfare. Again, this is a second-best result.

VII. Conclusion

In this paper we have developed a two-country overlapping-generations model of deterrence or conflict initiation. Our model emphasizes the intergenerational element of international tensions and so highlights aspects of arms accumulation that have received little attention in the literature. We abstract, however, from some topics that have received more attention in the literature. In particular, successive generations in our model play one-shot games, removing the scope for repeated-game phenomena, such as reputation effects and punishment strategies. We suspect that real-world behavior is characterized by repeated games with intergenerational externalities.

This paper follows in the tradition of the literature in that we examine the arms race in a two-country world. While such analysis was evidently appropriate during the Cold War era, it is less so today. Our approach remains useful for a number of reasons. First, the world may once again be dominated by two superpowers in the future. Second, and perhaps more importantly, our analysis also can be applied to regional conflicts when devastation of both countries is a possible outcome. Moreover, our model lends itself to a multicountry generalization with coalition formation; such an approach may provide a useful framework for analyzing present-day conflicts. Finally, a lesson of our model is that, although the world may no longer be dominated by two superpowers, the legacy of the Cold War affects current decisions and current welfare.

Despite our emphasis on intergenerational effects, some of our results resemble those of the arms-control literature with infinitely-lived agents. We find that arms-control (reduction) arrangements can either increase or decrease the possibility of international conflict and that international arrangements leading to increases in international weapons stocks are successful both in deterring conflict and in improving world welfare. While we do find that having no weapons is always better than having some weapons, fewer weapons need not be better, even if the reduction in weapons is brought about by an international agreement. Although apparently counterintuitive, this result is firmly grounded in intergenerational trade-offs: the immediate welfare improvements won by reducing the resources devoted to defense may be accompanied by uncompensated welfare losses on future generations. While the end of the Cold War may bring a peace dividend, it does not guarantee a safer world: current generations still bear the consequences of past accumulation decisions. This we see as a cautionary tale for the proponents of the "new world order."

REFERENCES

Allais, Maurice, *Economie et Intérêt*, Paris: Imprimerie Nationale, 1947.

Brams, Steven J. and Kilgour, D. Marc, "Optimal Deterrence," *Social Philosophy and Policy*, Autumn 1985, *3*, 118–35.

Cooper, Russell W. and John, A. Andrew, "Coordinating Coordination Failures in Keynesian Models," *Quarterly Journal of Economics*, August 1988, *103*, 441–64.

Garfinkel, Michelle R., "Arming as a Strategic Investment in a Cooperative Equilibrium," *American Economic Review*, March 1990, *80*, 50–68.

Intriligator, Michael D. and Brito, Dagobert L., "Can Arms Races Lead to the Outbreak of War?" *Journal of Conflict Resolution*, March 1984, *28*, 63–84.

John, A. Andrew, Pecchenino, Rowena A., Schimmelpfennig, David E. and Schreft, Stacey L., "Short-Lived Agents and the Long-Lived Environment," mimeo, Michigan State University, 1992.

Kotlikoff, Laurence J., Persson, Torsten and Svensson, Lars E. O., "Social Contracts as Assets: A Possible Solution to the Time Consistency Problem," *American Economic Review*, September 1988, *78*, 662–77.

Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*, 1002–37.

Samuelson, Paul A., "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *Journal of Political Economy*, December 1958, *66*, 467–82.

Sandler, Todd, "A Theory of Intergenerational Clubs," *Economic Inquiry*, April 1982, *20*, 191–208.

Shepherd, William, "Self Interest and National Security," *American Economic Review*, May 1988 (*Papers and Proceedings*), *78*, 50–4.

Slemrod, Joel B., "Saving and the Fear of Nuclear War," *Journal of Conflict Resolution*, September 1986, *30*, 403–19.

_____, "Fear of Nuclear War and Intercountry Differences in the Rate of Saving," *Economic Inquiry*, October 1990, *28*, 647–57.

Stokey, Nancy L., "Learning by Doing and the Introduction of New Goods," *Journal of Political Economy*, August 1988, *96*, 701–17.

van der Ploeg, Frederick and de Zeeuw, Aart J., "Perfect Equilibrium in a Model of Competitive Arms Accumulation," *International Economic Review*, February 1990, *31*, 131–46.

A Search-Theoretic Approach to Monetary Economics

By NOBUHIRO KIYOTAKI AND RANDALL WRIGHT*

The essential function of money is its role as a medium of exchange. We formalize this idea using a search-theoretic equilibrium model of the exchange process that captures the "double coincidence of wants problem" with pure barter. One advantage of the framework described here is that it is very tractable. We also show that the model can be used to address some substantive issues in monetary economics, including the potential welfare-enhancing role of money, the interaction between specialization and monetary exchange, and the possibility of equilibria with multiple fiat currencies. (*JEL* E00, D83)

Since the earliest writings of the classical economists it has been understood that the essential function of money is its role as a medium of exchange. The use of monetary exchange helps to overcome the difficulty associated with pure barter in economies where trade is not centralized through some perfect and frictionless market. Many attempts have been made in the literature to formalize this, with varying degrees of success.[1] In this paper, we present a search-theoretic equilibrium model of the exchange process that seems to capture the "double coincidence of wants problem" with pure barter in a simple and natural way. We show that this gives rise to a medium-of-exchange role for fiat currency. We also show that the model can be used to address some substantive issues in monetary economics.

Previously, in Kiyotaki and Wright (1989), we used a search-theoretic model to determine endogenously which commodities would become media of exchange, or commodity money. We were also able to construct an equilibrium with valued fiat currency; but since that model was designed primarily to study commodity money, it is not the most tractable framework within which to discuss fiat money. For instance, we only showed that fiat money could be valued if it has intrinsic properties at least as good as the best available commodity

*Kiyotaki: Federal Reserve Bank of Minneapolis and Department of Economics, University of Minnesota, Minneapolis, MN 55455; Wright: Federal Reserve Bank of Minneapolis and Department of Economics, University of Pennsylvania, Philadelphia, PA 19104. This is a much revised version of our earlier working paper, "Search for a Theory of Money." Part of the work on that project was accomplished while both authors were at the London School of Economics in the spring of 1990, and we are grateful for that institution's hospitality. The National Science Foundation, the University of Pennsylvania Research Foundation, and the Hoover Institution all provided financial support. We also thank many friends, colleagues, students, and seminar participants for their input. Peter Diamond, Robert Lucas, Dale Mortensen, David Romer, Neil Wallace, and two anonymous referees made some especially useful suggestions, although they should not be held accountable for the final product. The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

[1]There is a voluminous literature on the foundations of monetary theory, and rather than attempting to cite all of the relevant work, we refer the reader to the survey by Joseph M. Ostroy and Ross M. Starr (1990). We would, however, like to mention the contribution of Robert A. Jones (1976) and the extensions by Seongwhan Oh (1989) and Katsuhito Iwai (1988). Although there are many technical differences, that model is definitely related in spirit to the search-theoretic approach we describe here. In particular, there are heterogeneous agents and commodities, and in equilibrium certain commodities are chosen as media of exchange in order to reduce search costs.

money. In Kiyotaki and Wright (1991), we used an alternative search-based model to illustrate the robustness of monetary equilibria; that is, fiat money can be valued as a medium of exchange even if it has intrinsic properties, like its rate of return, that are inferior to other available assets. We also constructed an example in that model to show how the use of fiat money can affect welfare.

However, due to the generality of the specification in that paper, we were not able to say much about the features of monetary equilibria, other than that they exist and are robust, and our characterization of welfare did not proceed much beyond a numerical example. The model to be presented in this paper can be thought of as a simplified version of Kiyotaki and Wright (1991). Our first objective is to demonstrate that this class of models is actually very tractable. Our second objective is to convince the reader that search-based models can be used not just to determine which objects serve as media of exchange or to prove the existence of valued fiat money, but to address some more applied issues in monetary economics as well. In particular, we use the model to discuss the potential welfare-enhancing role of money, the interaction between specialization and monetary exchange, and the possibility of equilibria with multiple currencies.

The rest of the paper is organized as follows. In Section I we describe the basic model. In Section II we characterize the welfare effects of money. Among other things, the model implies that equilibria where fiat money is universally acceptable are generally superior to nonmonetary equilibria and to equilibria where it is only partially acceptable. In Section III we introduce specialization by producers, by assuming that they face a trade-off between productivity and the marketability of their output. The model implies that use of money, by making exchange easier, leads to more specialized and, therefore, more efficient production. In Section IV we discuss a version of the model that allows for multiple fiat currencies. In Section V we conclude. In order to improve the presentation we make some simplifying assumptions in specifying the basic model that one arguably may want to relax, and we show how to do so in the Appendix.

I. The Basic Model

The economy is populated by a large number of infinite-lived agents, with total population normalized to unity. There is also a large number of consumption goods. These consumption goods are indivisible and come in units of size one. We refer to them as *real commodities*, to distinguish them from *fiat money*, which is an object that no one ever consumes and can be thought of as a collection of pieces of paper or certain types of seashells, for example, with no intrinsic value. A crucial feature of the model is that there is an exogenous parameter x, with $0 < x < 1$, that captures the extent to which real commodities and tastes are differentiated. In particular, x equals the proportion of commodities that can be consumed by any given agent, and x also equals the proportion of agents that can consume any given commodity.[2] If a commodity is one of those that can be consumed by an agent, then we say that it is one of his consumption goods. Consuming one of his consumption goods yields utility $U > 0$, while consuming other commodities (or money) yields zero utility.

Initially, a fraction M of the agents are each endowed with money while $1 - M$ are each endowed with one real commodity, where $0 \leq M < 1$. Money may or may not have value. If it does, then it is convenient to assume that agents who are initially endowed with money are endowed with exactly one unit of real balances, so that in order to buy a real commodity they must spend all of their cash. There are two ways to guarantee that this is the case. First, and

[2] For example, suppose there are K distinct goods and each agent consumes k of them; then $x = k/K$. Alternatively, suppose there is a continuum of goods indexed by points around a circle of circumference 1 and each agent consumes goods corresponding to points in a fixed arc; then, x is the length of that arc.

most straightforwardly, we can simply assume that the monetary object is indivisible, like the real commodities in the model. Then if money trades at all it must trade one-for-one against a real commodity, and each agent with one indivisible unit of money will have one unit of real balances. Alternatively, we can assume that money is divisible, determine the price level endogenously for a given stock of nominal currency, and endow some agents at the initial date with exactly enough nominal currency to constitute a single unit of real balances. We begin with the former approach of assuming that money is indivisible and take up the latter, slightly more complicated, approach later.

Money and commodities are costlessly storable. Money cannot be produced by any private agent, while real commodities can be produced according to the following technology. One unit of output requires two inputs: a consumption good and a random amount of time. That is, once an agent consumes he enters a production process that yields one unit of one real commodity, drawn randomly from the set of all commodities, according to a continuous-time Poisson process with arrival rate $\alpha > 0$. Thus, α measures productivity in the sense of average output per unit time. Note that agents who have not consumed cannot produce. Furthermore, as is standard in the equilibrium search literature, we assume that agents cannot consume their own output (see e.g., Peter A. Diamond, 1982, 1984; Kiyotaki and Wright, 1991). This assumption helps to simplify the presentation and to facilitate comparison with earlier models, but as we show in Appendix A it is otherwise completely unnecessary.

An agent who has just produced enters an exchange sector where he looks for other agents with whom to trade. Traders in the exchange sector meet pairwise and at random according to a Poisson process with constant arrival rate $\beta > 0$.[3] When two traders meet, exchange takes place if and only if it is mutually agreeable, that is, if and only if both agents are at least as well off after the trade. Because there is a large number of anonymous agents, all trade is quid pro quo (there can be no IOU's or other forms of private credit). We also assume that there is a transaction cost ε in terms of disutility, where $0 < \varepsilon < U$, that must be paid by the receiver whenever any real commodity is accepted in trade. This transaction cost implies that a trader who is indifferent between holding two real commodities will never trade one for the other. For simplicity, we assume for most of the presentation that the transaction cost of accepting fiat money is zero; this simplifies the presentation considerably but, as we show in Appendix B, it is possible (and interesting) to relax this assumption.

Since exchange takes place if and only if mutually agreeable, an agent with either one unit of real balances or one real commodity cannot acquire additional money or another commodity except by giving up his entire inventory. Furthermore, no agent in the exchange sector can produce anything until he trades for one of his consumption goods and consumes, given the specified technology. These observations have the following implication: if each trader starts at the initial date with either one unit of real balances or one real commodity, then in equilibrium all traders will always have either one unit of real balances or one real commodity.[4] Agents with real commodities

[3]The assumption that the arrival rate β is constant (and independent of the number of traders) is equivalent to the assumption of a constant-returns-to-scale (CRS) meeting technology. That is, a CRS meeting technology implies that the total number of meetings per unit time is proportional to the number of traders, and so the arrival rate for a representative trader (which is just the number of meetings divided by the number of traders) is a fixed constant. We ignore degenerate outcomes in which there are no agents in the exchange sector, and hence, the arrival rate for an individual should he enter this sector would be 0.

[4]Note that there are no physical restrictions in the model against storing more than one commodity, storing arbitrary quantities of money, or storing money and commodities simultaneously. Rather, these results are due to the assumption that consumption is a necessary

are referred to as *commodity traders*, while agents with fiat money are referred to as *money traders*. Let μ denote the fraction of traders who are money traders, so that a trader located at random has money with probability μ and a real commodity with probability $1-\mu$.

Individuals choose strategies for deciding when to accept various commodities and fiat money in order to maximize their expected discounted utility from consumption net of transaction costs, taking as given the strategies of others. We look for Nash equilibria. We restrict attention for the most part to symmetric equilibria, where all agents and all real commodities are treated the same, and to steady-state equilibria, where strategies and all aggregate variables are constant over time. To construct the set of such equilibria, we describe some basic properties that they must satisfy, use these properties to describe an individual trader's best-response correspondence, and determine its fixed points.

The first thing to note is that an agent always accepts a real commodity if it is one of his consumption goods, whereupon he immediately consumes it and enters the production process. Also, we claim that a commodity trader will never accept a commodity that is not one of his consumption goods. This is due to the fact that in a symmetric equilibrium no real commodities are treated as special, and therefore, the probability of a trade offer being accepted by the next agent one meets is independent of the type of commodity one has. Hence, there is no advantage to trading one real commodity for another, and since there is a transaction cost ε, unless a commodity is going to be consumed it will never be accepted. This means that x is the probability that a commodity trader located at random is willing to accept any given commodity, and therefore x^2 is the probability that two commodity traders consummate a barter transaction. This is precisely William Stanley Jevons's (1875) "double coincidence of wants problem" with direct barter: not only do you have to meet someone with something that you want, this someone also has to want what you have.[5]

The next thing to determine is whether individuals accept money. Let Π denote the probability that a random commodity trader accepts money and let π be the best response of a representative individual. We will solve the best-response problem using dynamic programming. Let V_j denote the payoff or value function for the individual in state j, where $j=0$, 1, or m indicates that he is a producer, a commodity trader, or a money trader, respectively. Then, if $r>0$ is the rate of time preference, Bellman's equations are given by

$$(1)\quad rV_0=\alpha(V_1-V_0)$$

$$(2)\quad rV_1=\beta(1-\mu)x^2(U-\varepsilon+V_0-V_1)+\beta\mu x\max_{\pi}\pi(V_{\mathrm{m}}-V_1)$$

$$(3)\quad rV_{\mathrm{m}}=\beta(1-\mu)\Pi x(U-\varepsilon+V_0-V_{\mathrm{m}}).$$

input into production (an assumption we adopted from S. Rao Aiyagari and Neil Wallace [1991, 1992]) and the way we distribute the initial endowments. In principle, we could initially endow agents with more than one unit of commodities or real balances, but this would require solving for the steady-state inventory distribution and would lead to potentially complicated bargaining problems in bilateral exchange.

[5]The result that traders never accept commodities that are not their consumption goods means that there is no commodity money in a symmetric equilibrium. This is not to say that the model cannot have nonsymmetric equilibria, in which some real commodities do become media of exchange, but only that we restrict attention to symmetric outcomes here. Commodity money is analyzed in a related model in Kiyotaki and Wright (1989). Although a small transaction cost ε guarantees that there will be a double-coincidence problem in a symmetric equilibrium, the double-coincidence problem arises without transaction costs in the asymmetric equilibria studied in Kiyotaki and Wright (1989) and Aiyagari and Wallace (1991). Another way to guarantee that there is a double-coincidence problem is to assume that a real commodity can only be stored by its producer, as in Kiminori Matsuyama et al. (1993), which seems natural if we interpret these commodities as services rather than goods. Under this assumption, we can dispense with the transaction cost entirely.

Equations like these are standard in search theory (formal derivations for a closely related model can be found in Kiyotaki and Wright [1991], for example). They have the following interpretation. According to (1), the flow return to a producer, rV_0, equals the rate at which output is produced, α, times the gain from switching from production to exchange, $V_1 - V_0$. According to (2), the flow return to a commodity trader equals the sum of two terms. The first term is the rate at which he meets other commodity traders, $\beta(1-\mu)$, times the probability that both want to trade, x^2, times the gain from trading, consuming, and switching back to production, $U - \varepsilon + V_0 - V_1$. The second term is the rate at which he meets money traders, $\beta\mu$, times the probability that a money trader wants to trade, x, times the gain from accepting money with probability π, where π is chosen optimally. According to (3), the flow return to a money trader equals the rate at which he meets commodity traders, $\beta(1-\mu)$, times the probability that both want to trade, Πx, times the gain from trading, consuming, and switching to production, $U - \varepsilon + V_0 - V_m$.[6]

The above dynamic program depends not only on the strategies of others, as represented by Π, but also on μ, the proportion of traders holding money. However, μ can be determined as a function of Π and the initial endowment of money, M. Begin by letting N_0, N_1, and N_m denote the proportions of the population who are producers, commodity traders, and money traders. Then the model has a dynamic structure with transitions as illustrated in Figure 1. To determine its steady state, we equate the flow out of and into production:

$$(4)\quad \alpha N_0 = \beta(1-\mu)x^2 N_1 + \beta(1-\mu)\Pi x N_m.$$

If we use the fact that the N_j's sum to 1 and

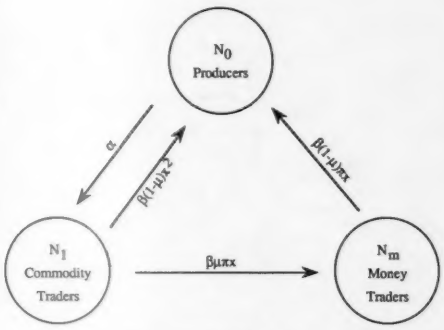


FIGURE 1. DYNAMIC STRUCTURE OF THE MODEL

the fact that $N_m = M$ (the number of money traders equals the number of agents endowed with money), (4) can be reduced to

$$(5)\quad M = \alpha\mu/(\alpha+\varphi)$$

where $\varphi = \varphi(\mu, \Pi)$ is defined by[7]

$$(6)\quad \varphi = \beta(1-\mu)\left[\mu x\Pi + (1-\mu)x^2\right].$$

Equation (5) is a quadratic in μ, and for any $M \in [0,1]$ and $\Pi \in [0,1]$ there will exist a unique value of $\mu = \mu(M,\Pi)$ in $[0,1]$ satisfying this equation. Furthermore, one can show that $\mu(0,\Pi) = 0$, $\mu(1,\Pi) = 1$, $\partial\mu/\partial M > 0$, and $\partial\mu/\partial\Pi > 0$. Given $\mu = \mu(M,\Pi)$, the unique steady state is fully described by

$$(7)\quad N_0 = \varphi/(\alpha+\varphi)$$

$$N_1 = (1-\mu)\alpha/(\alpha+\varphi).$$

However, for the purpose of analyzing the above dynamic program, μ summarizes all the agent needs to know about the steady state. If we insert $\mu = \mu(M,\Pi)$ into (1)–(3),

[6]We have implicitly assumed that a money trader never accepts a commodity that is not one of his consumption goods, but one can show that this is always true in equilibrium. That is, one can show that the only time an agent would want to exchange money for a commodity that is not one of his consumption goods is when money is valueless, in which case he cannot.

[7]Note that φ can be interpreted as consumption per trader per unit time: it is the rate at which a representative trader meets commodity traders, $\beta(1-\mu)$, times the probability that a deal is consummated, which is the probability our representative trader has money and they trade, $\mu x\Pi$, plus the probability that our representative trader has a real commodity and they trade, $(1-\mu)x^2$.

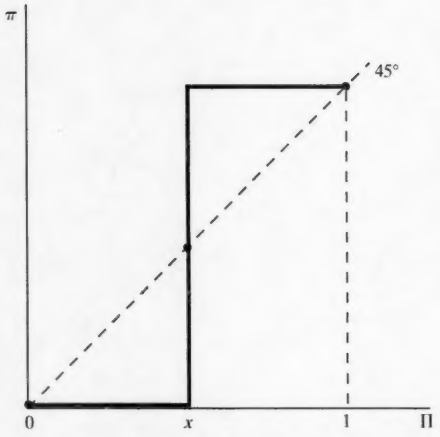


FIGURE 2. THE BEST-RESPONSE CORRESPONDENCE

then, given M, this dynamic program defines a correspondence from Π to best responses, π. The set of equilibria is the set of fixed points of this correspondence.

To characterize this set, first note that if $\Pi < x$ then (1)–(3) imply that $V_m < V_1$, which implies that the best response is $\pi = 0$. Intuitively, if money is being accepted with a lower probability than a barter offer, then it is harder to trade using money than barter, and so the best response is never to exchange a real commodity for money. Second, if $\Pi > x$, then (1)–(3) imply that $V_m > V_1$, which implies $\pi = 1$. If money is being accepted with a greater probability than a barter offer, then it is easier to trade using money, and so the best response is to exchange a real commodity for money whenever possible. Finally, if $\Pi = x$, then (1)–(3) imply that $V_m = V_1$, which implies that π can be anything in $[0,1]$. If monetary exchange and barter are equally easy then traders are indifferent between having money and real commodities, and they could accept money with any probability. Based on these results, the best-response correspondence is as shown in Figure 2, and there are exactly three equilibria: $\Pi = 0$, $\Pi = 1$, and $\Pi = x$.

The equilibrium with $\Pi = 0$ will be called the *nonmonetary equilibrium*. In this case, agents expect that money will be valueless, so they never accept it, and this expectation is self-fulfilling. The equilibrium with $\Pi = 1$ will be called the *pure-monetary equilibrium*. In this case, agents expect that money will be universally acceptable, and so they always take it, and this expectation is self-fulfilling. Finally, the equilibrium with $\Pi = x$ will be called the *mixed-monetary equilibrium*. In this case, agents are indifferent between accepting and rejecting money as long as other agents take it with probability $\Pi = x$, and so partial acceptability can also be self-fulfilling. Alternatively, a symmetric mixed-strategy equilibrium where all agents accept money with probability x could be reinterpreted as a nonsymmetric pure-strategy equilibrium, where a fraction x of agents accept money with probability 1 while the rest accept it with probability 0.[8]

II. Welfare

The first thing we want to do in this section is to compare utility across the various equilibria, for a given value of M. For the purpose of this comparison, we keep things tractable by restricting attention to the limiting case where $\alpha \to \infty$. In this case production is instantaneous, and so all agents are either money traders or commodity traders: $N_m = M$, $N_1 = 1 - M$, and $\mu = M$. This makes it relatively easy to solve (1)–(3) for the reduced-form payoffs:

$$(8) \quad rV_1 = \psi\{rx + \beta x \Pi[M\Pi + (1-M)x]\}$$

$$(9) \quad rV_m = \psi\{r\Pi + \beta x \Pi[M\Pi + (1-M)x]\}$$

where $\psi = (U - \varepsilon)\beta(1 - M)x/(r + \beta x \Pi)$. We can now substitute $\Pi = 0$, $\Pi = x$, and $\Pi = 1$ into (8) and (9) and compare utility across equilibria for commodity traders and money traders.

[8]There can exist non-steady-state equilibria in this model where the probability that money is accepted varies over time. An example of a "sunspot equilibrium," in which the probability that money is accepted fluctuates randomly over time even though the fundamentals are nonstochastic and time-invariant, is constructed in Kiyotaki and Wright (1990).

If we let the superscripts N, M, and P refer to the nonmonetary, mixed-monetary, and pure-monetary equilibria, respectively, then we have the following results. First, commodity traders are equally well off in the nonmonetary and mixed-monetary equilibria and strictly better off in the pure-monetary equilibrium: $V_1^N = V_1^M < V_1^P$. Second, money traders are strictly better off in the pure-monetary equilibrium than in the mixed-monetary equilibrium and strictly better off in the mixed-monetary equilibrium than in the nonmonetary equilibrium: $V_m^N < V_m^M < V_m^P$. Thus, given the initial endowment of money and real commodities, all agents are at least weakly better off if money is acceptable than if it is not, and all agents are strictly better off if money is universally acceptable than if it is only partially acceptable.[9]

The next thing we want to do is to examine how utility varies with M, and for the purpose of this comparison we return to the general case where α need not be ∞. The experiment we consider is to increase the number of agents initially endowed with money and to reduce the number initially endowed with real output, so that we can maintain the tractability of the unit-inventory assumption. In either the nonmonetary equilibrium or the mixed-monetary equilibrium, all agents are better off the lower is M. The reason is that, in these equilibria, money does nothing to ameliorate the double-coincidence problem, and so it is better to endow everyone with real consumption goods rather than intrinsically worthless paper or seashells. The more interesting case is the pure-monetary equilibrium, where fiat currency does have a genuine role to play in facilitating exchange.

[9]These results differ from those in Kiyotaki and Wright (1990), where we assumed that agents initially endowed with fiat currency would freely dispose of it and produce a new commodity in the nonmonetary equilibrium. This made the initial stock of real commodities differ across monetary and nonmonetary equilibria and therefore made welfare comparisons ambiguous. Following Aiyagari and Wallace (1992), we assume here that agents initially endowed with fiat money cannot produce until they consume, which keeps the initial stock of output constant across equilibria.

To pursue this, define the welfare criterion

$$W = N_0 V_0 + N_1 V_1 + N_m V_m. \tag{10}$$

This can be interpreted as the *ex ante* expected utility of all agents before the initial endowment of money and output is randomly distributed among them. After some algebra, (10) can be simplified to yield

$$rW = (U - \varepsilon)\varphi\alpha/(\alpha + \varphi) \tag{11}$$

where $\varphi = \varphi(\mu, \Pi)$ was defined in (6) above.[10] Now consider maximizing W with respect to M. Since W is increasing in φ, we proceed by finding the value μ^o that maximizes φ with respect to μ, and then determine the optimal value M^o from the steady-state condition (5), $M = \alpha\mu/(\alpha + \varphi)$.

The result is as follows: if $x \geq \frac{1}{2}$ then $\mu^o = 0$, which implies $M^o = 0$; if $x < \frac{1}{2}$ then $\mu^o = (1-2x)/(2-2x)$, which implies $M^o > 0$. Intuitively, when $x \geq \frac{1}{2}$, each agent is willing to consume (and therefore accept) at least half of the commodities produced in the economy, and pure barter is not very difficult. In this case the role for a medium of exchange is not very important, and it is optimal to endow everyone with real output and no one with money. When $x < \frac{1}{2}$, on the other hand, pure barter is sufficiently difficult that the introduction of some fiat money improves welfare, in spite of the fact that, in the experiment under consideration, endowing some agents with money requires endowing fewer agents with real output at the initial date. We also note that M^o is decreasing in x, and that $M^o \to \frac{1}{2}$ as $x \to 0$. Thus, as x shrinks and the double-coincidence problem becomes more difficult it is optimal to endow more agents with money.

We now turn to a version of the model in which money can be interpreted as being divisible rather than indivisible and investigate the welfare implications of a particular mechanism for determining the price level,

[10]Notice that rW equals the difference $U - \varepsilon$ times aggregate consumption, since φ is consumption per trader per unit time and $\alpha/(\alpha + \varphi) = N_1 + N_m$ is the number of traders.

P. We look for a pure-monetary equilibrium in which each money trader carries P units of cash and all P units are required to purchase one real commodity. Then real balances are given by $M = C/P$ for any exogenous stock of nominal currency C. Of course, to make P endogenous we need to impose an additional equilibrium condition. Consider the method used by Diamond (1984) in his cash-in-advance search model, which is to impose as an equilibrium condition that the gains from trade for a commodity trader and a money trader are equalized whenever an exchange is made between them:

$$(12) \qquad V_m - V_1 = U - \varepsilon + V_0 - V_m.$$

The left-hand side is the gain from trade for a commodity trader who accepts money, while the right-hand side is the gain for a money trader who acquires one of his consumption goods.

If a pure-monetary equilibrium satisfies condition (12), we call it a *split-the-surplus* equilibrium. Notice that both sides of (12) depend on μ. If μ is large, there are many money traders and few commodity traders, so having money is not very desirable; hence, a commodity trader who acquires money gets a smaller gain than a money trader who acquires one of his consumption goods. Thus, for large μ the right-hand side of (12) exceeds the left-hand side, and we need to reduce μ until either (12) holds or we hit $\mu = 0$. Inserting the reduced-form payoff functions and simplifying, the unique value of μ that satisfies (12) is given by

$$(13) \qquad \mu^* = (1-2x)/(2-2x) - r/2\beta x(1-x).$$

If $r < \beta x(1-2x)$ then $\mu^* > 0$, which implies a unique $M^* > 0$ satisfying (5), and a finite equilibrium price level $P^* = C/M^*$. If $r \geq \beta x(1-2x)$, then (12) cannot be satisfied for any value of $\mu > 0$. In this case, we say that the split-the-surplus equilibrium entails $M^* = 0$ and $P = \infty$, and hence there can be no monetary exchange.

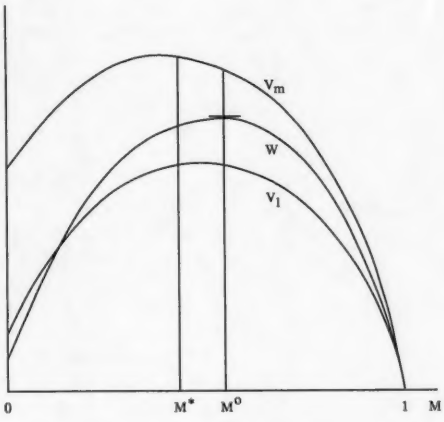


FIGURE 3. VALUE FUNCTIONS AND WELFARE

Recall that the value of μ that maximizes W is $\mu^o = (1-2x)/(2-2x)$, and comparing this with (13) we find $\mu^o > \mu^*$. This means the split-the-surplus equilibrium yields a lower value of μ, and hence a lower value of M and a higher value of P for any given C, than that which maximizes *ex ante* utility. However, the split-the-surplus equilibrium is still *ex post* Pareto optimal. To see why, consider the value functions V_m and V_1 (V_0 is proportional to V_1 and need not be considered independently). One can show that, as functions of M, both are concave, and V_1 is increasing but V_m is decreasing at $M = M^*$ (see Fig. 3). Hence, any movement away from M^* that makes commodity traders better off makes money traders worse off, and vice versa. The split-the-surplus equilibrium is therefore *ex post* efficient even though it fails to maximize *ex ante* welfare.[11]

[11] Figure 3 indicates that both commodity and money traders prefer a lower value of M than that which maximizes W. This ostensibly paradoxical result can be understood by noting that the number of agents of each type varies with M.

III. Specialization

An insight dating back at least to Adam Smith (1776) is that specialization is limited by the extent of the market and that the use of money encourages specialization by enlarging the extent of the market. As Smith puts it:

> When the division of labour has been once thoroughly established, it is but a very small part of a man's wants which the produce of his own labour can supply. He supplies the far greater part of them by exchanging that surplus part of the produce of his own labour, which is over and above his own consumption, for such parts of the produce of other men's labour as he has occasion for. Every man thus lives by exchanging, or becomes in some measure a merchant, and the society itself grows to be what is properly a commercial society.
>
> But when the division of labour first began to take place, this power of exchanging must frequently have been very much clogged and embarrassed in its operations. One man, we shall suppose, has more of a certain commodity than he himself has occasion for, while another has less. ... But if this latter should chance to have nothing that the former stands in need of, no exchange can be made between them. ... In order to avoid the inconveniency of such situations, every prudent man in every period of society, after the first establishment of the division of labour, must naturally have endeavored to manage his affairs in such a manner, as to have at all times by him, besides the peculiar produce of his own industry, a certain quantity of some one commodity or other, such as he imagined few people would be likely to refuse in exchange for the produce of their industry.
>
> [1937 pp. 22–3]

Smith is suggesting that specialization, while it may have desirable consequences in terms of productivity, makes barter difficult. Whenever they can, specialized producers will therefore tend to sell their output for a more generally acceptable medium of exchange, which they will then use to buy whatever consumption goods they desire. Hence, specialization leads to a greater role for money, while at the same time the use of money affords a greater opportunity for specialization by facilitating the process of exchange. In order to formalize this, we introduce a trade-off between productivity and marketability by assuming that the arrival rate in the production process is a function of the number of agents willing to consume the output: $\alpha = \alpha(x)$, where $\alpha' < 0$. The idea is that, by becoming more specialized, a producer can increase output per unit time, α, but only at the cost of reducing the fraction of consumers who will accept his output in exchange, x.[12]

Before entering the production process, agents choose x, taking as given the behavior of others. If money is accepted with probability Π and other producers' decisions imply that a given individual can consume a fraction X of their output, his payoff if he chooses x is described by

$$(14)\quad rV_0 = \alpha(x)[V_1(x) - V_0]$$

$$(15)\quad rV_1(x) = \beta(1-\mu)Xx[U - \varepsilon + V_0 - V_1(x)] + \beta\mu x\Pi[V_m - V_1(x)]$$

$$(16)\quad rV_m = \beta(1-\mu)X\Pi(U - \varepsilon + V_0 - V_m).$$

The choice of x will be made by a producer to maximize the right-hand side of (14), and this x is then carried over to the exchange sector as a state variable. As was the case earlier, this best-response problem also depends on μ, but μ will be determined below

[12] One interpretation is that each consumer derives utility from a fixed set of *characteristics* embodied in some commodities, and a larger value of x implies that the producer's output contains a greater number of characteristics and hence has a larger potential market. Eduardo Siandra (1990) has independently developed a very similar model. Robert King and Charles Plosser (1986) and Harold L. Cole and Alan C. Stockman (1992) provide other analyses of the interaction between money and specialization.

using the steady-state conditions as a function of strategies and M.

Equations (14)–(16) can be solved for V_0, V_1, and V_m. In particular, after simplification, we find that

$$(17)\quad \alpha(x)[V_1(x)-V_0] = \phi x\alpha(x)/\zeta(x) \equiv Z(x)$$

where

$$\zeta(x) = [r+\beta(1-\mu)X\Pi][r+\beta(1-\mu)xX+\beta\mu x\Pi] + \alpha(x)[r+\beta(1-\mu)X\Pi+\beta\mu x\Pi]$$

and ϕ does not depend on x. The individual choice of x can be found by maximizing $Z(x)$. If we assume an interior solution, the first-order condition $Z'(x)=0$ can be rearranged to yield:[13]

$$(18)\quad -\frac{x\alpha'(x)}{\alpha(x)} = \frac{r+\alpha(x)}{r+\beta(1-\mu)Xx+\beta\mu x\Pi}.$$

It can be shown that the second-order condition $Z''<0$ holds if we assume that $\alpha''\le 0$. Then (18) completely characterizes the individual's choice of x, given X, Π, and μ. We write $x = x(X,\mu,\Pi)$.

For a symmetric equilibrium we must have $X=x$, or $X=x(X,\mu,\Pi)$. Another equilibrium condition comes from the steady-state equation $M=\alpha\mu/(\alpha+\varphi)$. Since the right-hand side of this equation depends on X, μ, and Π, we write it as $M=M(X,\mu,\Pi)$. Given M, an equilibrium is a solution to $X=x(X,\mu,\Pi)$ and $M=M(X,\mu,\Pi)$ with either $\Pi=0$, $\Pi=X$, or $\Pi=1$ (since, for any given X, the model has a nonmonetary equilibrium, a mixed-monetary equilibrium, and a pure-monetary equilibrium, exactly as in the model without endogenous specialization). In Figure 4 we draw the locus of points in (μ,X)-space satisfying each of these conditions. Notice that the $M=M(X,\mu,\Pi)$ curve is upward-sloping, shifts to the right as Π increases, and goes through $(M,0)$. Also, the $X=x(X,\mu,\Pi)$ curve is upward-sloping, horizontal, or downward-sloping depending on whether $\Pi=0$, X, or 1, and goes through the same intercept $(0,X_0)$ in any case. The intersection of these two curves determines the equilibrium values of μ and X in the nonmonetary equilibrium, the mixed-monetary equilibrium, or the pure-monetary equilibrium, depending on whether $\Pi=0$, $\Pi=X$, or $\Pi=1$.[14]

Again let the superscripts N, M, and P represent the nonmonetary, mixed-monetary, and pure-monetary equilibria. As can be seen from the diagram, specialization is greatest in the pure-monetary equilibrium, lower in the mixed-monetary equilibrium, and lowest in the nonmonetary equilibrium: $X^P < X^M < X^N$. The intuition behind this result is that when money circulates there is less of an advantage to having a high value of x, since it does not necessarily require a double coincidence of wants in order to exchange. We can also ask how specialization depends on M. An increase in M shifts the $M=M(X,\mu,\Pi)$ curves to the right but does not affect the $X=x(X,\mu,\Pi)$ curves. As can be seen from the diagram, when M increases, the result is an increase in X in the nonmonetary equilibrium, a decrease in

[13] One can also maximize the right-hand side of (14) directly by setting

$$\alpha'(x)[V_1(x)-V_0]+\alpha(x)V_1'(x)=0$$

where $V_1'(x)$ is, from (15),

$$V_1'(x) = rV_1(x)\Big/\{rx+\beta x^2[(1-\mu)X+\mu\Pi]\}.$$

Manipulating these equations yields the same first-order condition as in the text, equation (18).

[14] Notice that the pure- and mixed-monetary equilibria must be unique, but since $\Pi=0$ implies that both the $X=x(X,\mu,\Pi)$ curve and the $M=M(X,\mu,\Pi)$ curve are upward-sloping, they could intersect more than once, and there could be more than one nonmonetary equilibrium. Although examples with multiple nonmonetary equilibria can be constructed, we rule this out in the following discussion.

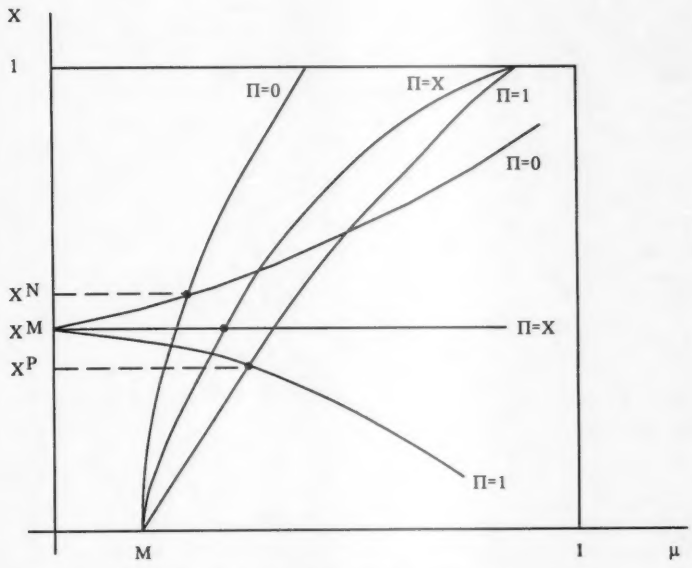


FIGURE 4. M AND X CURVES, WHOSE INTERSECTIONS DETERMINE EQUILIBRIUM VALUES OF μ AND X

X in the pure-monetary equilibrium, and no change in the mixed-monetary equilibrium. Roughly speaking, an increase in M in the pure-monetary equilibrium encourages specialization because producers can more easily market their specialized output when there is more money in circulation.[15]

Consider now the effect of increasing the arrival rate in the exchange sector, β, which can be thought of as reducing the frictions associated with trade (or "increasing the extent of the market"). In any pure-monetary equilibrium, an increase in β shifts the $X = x(X, \mu, 1)$ curve down and shifts the $M = M(X, \mu, 1)$ curve to the right, resulting in a decline in X and therefore an increase in specialization and productivity. As $\beta \to \infty$, $x \to 0$, and specialization becomes complete. As this happens, barter becomes extremely difficult, and the ratio of the volume of barter to monetary exchange vanishes.[16] In the limit, agents almost always sell their production goods for money and use money to buy their consumption goods; as Robert W. Clower (1965) puts it, "money buys goods and goods buy money; but goods do not buy goods." In this model, however, there is no constraint that agents *must* use cash. To

[15]The discussion of the effect of an increase in M takes the real money supply to be exogenous, say, because the monetary object is indivisible. Alternatively, we can assume that money is divisible and determine the level of real balances endogenously in pure-monetary equilibrium, given nominal balances, using the split-the-surplus condition discussed above. In terms of Figure 4, we need to shift the $M = M(X, \mu, 1)$ curve until the gains from trade for commodity traders and money traders are equalized. One can show that there exists a unique split-the-surplus equilibrium, and it implies a finite price level under appropriate parameter restrictions, as in Section II.

[16]The rate of barter exchange is $\beta(1-\mu)^2x^2$, while the rate of monetary exchange is $\beta\mu(1-\mu)x$. The ratio of these two is $(1-\mu)x/\mu$, which vanishes as $x \to 0$ (note that μ is bounded below by M). For any finite β, however, there will always be *some* direct barter in equilibrium.

the contrary, it is because the economy has settled on the use of a generally acceptable currency that specialization becomes profitable, and it is specialization that inhibits barter.

IV. Dual Currency Regimes

In this section we take up the possibility of multiple fiat monies. It is motivated by the observation that, in some economies, there seems to be more than one type of currency in simultaneous circulation. For instance, it is possible in certain locations to have both a domestic currency and a foreign currency used in exchange, although perhaps the former is generally acceptable while the latter is only partially acceptable. One example is that Canadian dollars are often accepted just across the U.S. border, and vice versa, although the foreign currencies are not always accepted by domestic residents. Furthermore, this situation can persist even if the two currencies differ in terms of rates of return or other intrinsic properties.

In order to study the phenomenon of dual-currency equilibria, we assume that there are now two colors of fiat money: red and blue. To simplify the presentation as much as possible, we only consider the case in which specialization is exogenous, and we assume that both monies are indivisible. If we endow all agents with either one unit of red money, one unit of blue money, or one real commodity at the initial date, then all agents will always hold one and only one of these objects at all future dates as well. We give the monies potentially different intrinsic properties by letting y_R and y_B denote flow yields or dividends; that is, each money yields y_j "utils" to its bearer per unit time (if $y_j < 0$ then it can be thought of as a storage cost). Also, let the supplies of the two monies be M_R and M_B, with $M_R + M_B < 1$, let μ_R and μ_B be the proportions of traders with red money and blue money, and let $\mu_C = 1 - \mu_R - \mu_B$ be the proportion of traders with real commodities.

To formulate the representative individual's best-response problem, let the probabilities of random commodity traders accepting red money and blue money be Π_R and Π_B. Then Bellman's equations are described by

$$(19)\quad rV_0 = \alpha(V_1 - V_0)$$

$$(20)\quad rV_1 = \beta\mu_C x^2(U - \varepsilon + V_0 - V_1) + \beta\mu_R x \max_{\pi_R} \pi_R(V_R - V_1) + \beta\mu_B x \max_{\pi_B} \pi_B(V_B - V_1)$$

$$(21)\quad rV_R = y_R + \beta\mu_C x\Pi_R(U - \varepsilon + V_0 - V_R)$$

$$(22)\quad rV_B = y_B + \beta\mu_C x\Pi_B(U - \varepsilon + V_0 - V_B).$$

These depend on μ_R and μ_B, but as above, the steady-state conditions can be solved for unique values of μ_R and μ_B, given strategies and exogenous values of M_R and M_B.[17]

Our goal is to construct an equilibrium in which both monies circulate, but with different acceptabilities: $1 = \Pi_R > \Pi_B > 0$. This requires $V_R > V_1 = V_B$. Now, $V_R > V_1$ follows immediately from $\Pi_R = 1$. Furthermore, for the case in which $y_R = y_B = 0$, (20)–(22) imply that $V_B = V_1$ if and only if

$$(23)\qquad \Pi_B = \Delta x$$

where $\Delta = (r + \beta x\mu_C + \beta\mu_R)/(r + \beta x\mu_C + \beta x\mu_R)$. Notice that $\Delta > 1$. If $\Pi_R = 1$ and $\Pi_B = \Delta x$, we have an equilibrium in which red money is universally accepted while blue money is only partially accepted. By continuity, we can perturb y_R and y_B without destroying the equilibrium, as long as $|y_R|$ and $|y_B|$ are not too great. In particular, we can construct equilibria with $1 = \Pi_R > \Pi_B$ even though $y_R < y_B$. In such an equilibrium, both monies circulate, but the high-return asset is less acceptable or less liquid

[17]The way we write Bellman's equations implicitly assumes that agents never trade one currency for another, which is true in equilibrium because such a trade could not possibly make both agents better off.

than the low-return asset. That is, the reason why red money is universally acceptable, even though it is dominated in rate of return, is that it has liquidity value. If the spread $y_B - y_R$ becomes too big, however, this equilibrium can no longer exist.[18]

VI. Conclusion

We have presented a model of exchange in which the difficulty of pure barter leads to a transactions role for fiat currency, and we have used the model to address several issues in monetary economics.[19] Other applications can also be studied in this framework. In Matsuyama et al. (1993), an explicit two-country version of a model similar to the one presented here is considered and used to investigate some issues relating to international monetary theory. In Steve Williamson and Wright (1991), a "lemons" problem is introduced into the model and used to illustrate the role of fiat currency in helping to overcome the frictions associated with private information. Siandra (1990) considers further the relationship between specialization and monetary exchange. Victor E. Li (1991) pursues some issues relating to externalities, welfare, and policy. Ramon Marimon et al. (1990) use a related model to analyze learning. Aiyagari and Wallace (1992) consider several other applications. Although there are many unanswered questions and much work remains to be done, we think that these search-theoretic models have definitely enhanced our understanding of the exchange process, in general, and of money, in particular.

[18]Similar arguments can be used to show that there are equilibria in which both monies are universally acceptable even though one has a higher rate of return, and equilibria in which one money circulates but the other does not even though one or the other has a higher rate of return. Examples with two circulating fiat currencies have also been constructed by Aiyagari and Wallace (1992).

[19]In order to focus on more substantive issues, we have neglected many of the technical aspects of search-based exchange models, like the possibility of multiple dynamic equilibria, including sunspot and cyclical equilibria (see Kiyotaki and Wright, 1990; Michele Boldrin et al., 1991; Timothy J. Kehoe et al. 1991).

Appendix A

Here we sketch a version of the model that makes it precise why agents need to trade, without assuming that they cannot consume their own output. This is perhaps more satisfying, but it does entail an increase in notation. The implications of this version of the model are essentially the same as those described in the text. For simplicity, we consider only the case where $\alpha = \infty$, but it should be clear how to handle the more general model.

Suppose there are K types of agents, with equal numbers of each type, and K commodities, where $K \geq 3$. Agents are specialists in production but generalists in consumption, in the following sense. Each agent can produce only some of the commodities —to ease the presentation, suppose each type can produce exactly one commodity— but has a need to consume different things at different points in time. In particular, after consuming one commodity a consumer realizes a taste or need for a new commodity drawn at random. That is, the probability that the new commodity will be j is $1/K$ for any $j = 1, 2, \ldots, K$. A consumer with a need for commodity j gets utility U from consuming it and no utility from anything else, at least until j is consumed and a new taste shock is realized.

Consider a representative agent. After consumption, with probability $1/K$ he needs the commodity he can produce, consumes immediately, and draws a new taste shock, while with probability $(K-1)/K$ he needs something else and must attempt to acquire it through trade. Therefore, the expected value of drawing a new taste shock, say V_n, will satisfy $V_n = (U + V_n)/K + V_1(K-1)/K$. Eventually, our representative agent needs something that he cannot produce. When he meets a potential trading partner, the probability that this partner needs the good our representative agent produces is $1/K$, while the probability that this partner also has the good our representative agent needs is $1/(K-1)$, since he must have one of the commodities other than the one that he himself needs. Hence, the probability of a double coincidence is $1/K(K-1)$. This im-

plies that the value functions for commodity and money traders satisfy

$$rV_1 = \frac{\beta(1-M)}{K(K-1)}(U-\varepsilon+V_n-V_1) + \frac{\beta M}{K}\pi(V_m-V_1)$$

$$rV_m = \frac{\beta(1-M)}{K}\Pi(U-\varepsilon+V_n-V_m).$$

The rest of the analysis is the same as the model in the text.

Appendix B

Here we show how to relax the assumption of a zero transaction cost on accepting money, where, for simplicity, we again consider only the case where $\alpha=\infty$. Let η denote the disutility cost of accepting currency, and assume that $0 \le \eta < U-\varepsilon$. The value functions in this generalized model satisfy

$$rV_1 = \beta(1-M)x^2(U-\varepsilon) + \beta Mx\pi(V_m-V_1-\eta)$$

$$rV_m = \beta(1-M)x\Pi(U-\varepsilon+V_1-V_m).$$

The only difference from the model in the text is that the gain to accepting money is now $V_m-V_1-\eta$. As before, there is always a nonmonetary equilibrium with $\Pi=0$. There is a pure-monetary equilibrium if and only if $V_m-V_1-\eta \ge 0$ when evaluated at $\Pi=1$. Manipulating the above equations implies that this is true if and only if $\eta \le \overline{\eta}$, where

$$\overline{\eta} \equiv \frac{\beta(1-M)x(1-x)(U-\varepsilon)}{r+\beta x(1-M)}.$$

Finally, there is a mixed-monetary equilibrium if and only if $V_m-V_1-\eta=0$ when evaluated at $\Pi=\Pi^M \in (0,1)$. Manipulating the above equations implies that $V_m-V_1-\eta=0$ if and only if

$$\Pi^M = \frac{\beta(1-M)x^2(U-\varepsilon)+r\eta}{\beta(1-M)x(U-\varepsilon-\eta)}.$$

Notice that $\Pi^M>0$ and $\Pi^M<1$ if and only if $\eta<\overline{\eta}$.

Hence, there exists the same set of three equilibria for any η in $(0,\overline{\eta})$. In terms of Figure 2, an increase in η shifts the best-response correspondence to the right. For $\eta<\overline{\eta}$ there are still three intersections with the 45° line. For $\eta>\overline{\eta}$ there is only one, at $\Pi=0$. Since $\overline{\eta}>0$, we can accommodate a positive transaction cost on money. Furthermore, $\overline{\eta}$ can exceed ε (for example, as long as U is sufficiently large), and we can even accommodate a transaction cost on money that exceeds the transaction cost on real commodities. We can also use the split-the-surplus condition to determine M^* and $P^*=M^*/C$ endogenously for any given stock of nominal balances C. It can be shown that $M^*>0$, and therefore $P^*<\infty$, if and only if $\beta x(1-2x)-r>\eta(U-\varepsilon)/(r+\beta x)$, which generalizes the condition for a finite price level given immediately after (13).

REFERENCES

Aiyagari, S. Rao and Wallace, Neil, "Existence of Active Trade Steady States in the Kiyotaki-Wright Model," *Review of Economic Studies*, October 1991, *58*, 901–16.

_____ **and** _____, "Fiat Money in the Kiyotaki-Wright Model," *Economic Theory*, 1992, *2* (4), 447–64.

Boldrin, Michele, Kiyotaki, Nobuhiro and Wright, Randall, "A Dynamic Equilibrium Model of Search, Production and Exchange," Northwestern University CMSEMS Discussion Paper No. 930, 1991; *Journal of Economic Dynamics and Control* (forthcoming).

Clower, Robert W., "A Reconsideration of the Microfoundations of Monetary Theory," *Western Economic Journal*, December 1965, *6*, 1–8.

Cole, Harold L. and Stockman, Alan C., "Specialization, Transactions Technologies,

and Money Growth," *International Economic Review*, May 1992, *33*, 283–98.

Diamond, Peter A., "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, *90*, 881–94.

_____, "Money in Search Equilibrium," *Econometrica*, January 1984, *52*, 1–20.

Iwai, Katsuhito, "The Evolution of Money: A Search Theoretic Foundation of Monetary Economics," University of Pennsylvania CARESS Working Paper No. 88-03, 1988.

Jevons, William Stanley, *Money and the Mechanism of Exchange*, London: Appleton, 1875.

Jones, Robert A., "The Origin and Development of Media of Exchange," *Journal of Political Economy*, August 1976, *84*, 757–75.

Kehoe, Timothy J., Kiyotaki, Nobuhiro and Wright, Randall, "More on Money as a Medium of Exchange," Federal Reserve Bank of Minneapolis Staff Report 140, 1991; *Economic Theory* (forthcoming).

King, Robert G. and Plosser, Charles I., "Money as the Mechanism of Exchange," *Journal of Monetary Economics*, January 1986, *17*, 93–115.

Kiyotaki, Nobuhiro and Wright, Randall, "On Money as a Medium of Exchange," *Journal of Political Economy*, August 1989, *97*, 927–54.

_____ **and** _____, "Search for a Theory of Money," National Bureau of Economic Research (Cambridge, MA) Working Paper No. 3482, 1990.

_____ **and** _____, "A Contribution to the Pure Theory of Money," *Journal of Economic Theory*, April 1991, *53*, 215–35.

Li, Victor E., "Money in Decentralized Exchange and the Optimal Rate of Inflation," mimeo, Northwestern University, November 1991.

Marimon, Ramon, McGrattan, Ellen and Sargent, Thomas J., "Money as a Medium of Exchange in an Economy with Artificially Intelligent Agents," *Journal of Economic Dynamics and Control*, May 1990, *14*, 329–74.

Matsuyama, Kiminori, Kiyotaki, Nobuhiro and Matsui, Akihiko, "Toward a Theory of International Currency," *Review of Economic Studies*, 1993 (forthcoming).

Oh, Seongwhan, "A Theory of a Generally Acceptable Medium of Exchange and Barter," *Journal of Monetary Economics*, January 1989, *23*, 101–19.

Ostroy, Joseph M. and Starr, Ross M., "The Transactions Role of Money," in Benjamin M. Friedman and Frank K. Hahn, eds., *Handbook of Monetary Economics*, Amsterdam: North-Holland, 1990, pp. 3–62.

Siandra, Eduardo, "Money and Specialization in Production," Department of Economics Working Paper No. 610, University of California, Los Angeles, 1990.

Smith, Adam, *An Inquiry into the Nature and Causes of the Wealth of Nations*, London: W. Strahan and T. Cadell, 1776; reprinted, New York: Modern Library, 1937.

Williamson, Steve and Wright, Randall, "Barter and Monetary Exchange Under Private Information," Federal Reserve Bank of Minneapolis Staff Report 141, 1991.

Monetary Policy and Credit Conditions: Evidence from the Composition of External Finance

By ANIL K KASHYAP, JEREMY C. STEIN, AND DAVID W. WILCOX*

In this paper, we use the relative moments in bank loans and commercial paper to provide evidence on the existence of a loan-supply channel of monetary-policy transmission. We find that tighter monetary policy leads to a shift in firms' mix of external financing: commercial paper issuance rises while bank loans fall. This suggests that contractionary policy can indeed reduce loan supply. Furthermore, such shifts in loan supply seem to affect investment, even controlling for interest rates and output. (*JEL* E44, E52, G32)

How does monetary policy affect the economy? The textbook *IS-LM* model holds that monetary policy operates through the liability side of banks' balance sheets: when the Fed tightens policy by draining reserves from the banking system, the noteworthy consequence is a fall in the stock of money. An alternative view is that independent effects come from the asset side of banks' balance sheets (i.e., from bank loans). According to this view, there are some borrowers for whom nonbank sources of credit do not represent a perfect substitute for bank loans. Therefore, when tight monetary policy shrinks the size of the banking sector, it reduces the overall supply of loans to these "bank-dependent" borrowers. As a result, investment and aggregate demand fall by more than can be accounted for by the conventional money channel.[1]

A substantial body of empirical work seeks to distinguish between the "money" and "lending" theories of the transmission mechanism. Much of this work focuses on comparing the correlation between output and money to the correlation between output and loans.[2] This approach has its limitations, though, since it does not explicitly deal with the issue of endogeneity.[3] For example, a finding that output is more closely correlated with money than with loans does not necessarily mean that money is more important in a causal sense. It could instead be that money itself is endogenous and responds more strongly than do loans to exogenous output shocks.

Two recent papers revisit this issue. Bernanke and Blinder (1992) and Christina Romer and David Romer (1990) both identify what are arguably good indicators of exogenous shifts in Federal Reserve policy.

*Kashyap: Graduate School of Business, University of Chicago, Chicago, IL 60637, and National Bureau of Economic Research; Stein: MIT-Sloan School of Management, Cambridge, MA 02139, and National Bureau of Economic Research; Wilcox: Federal Reserve Board, Washington, DC 20551. The authors are grateful to Ben Bernanke, William English, Ben Friedman, Dale Henderson, Eileen Mauskopf, Andrew Rose, Jeff Wooldridge, and seminar participants at various institutions for helpful comments and suggestions, and to Steve Helwig and Peter Simon for expert research assistance. Thanks also to the Federal Reserve Bank of Chicago, the National Science Foundation, the International Financial Services Research Center, and Batterymarch Financial Management for financial and research support. The views in this paper are solely those of the authors and do not reflect the opinions of the Board of Governors of the Federal Reserve System or its staff.

[1] The idea that there may be an important lending channel has a long history. Early work on the subject includes James Tobin and William Brainard (1963) and Brainard (1964). For a recent exposition of the distinction between the money and lending channels of policy transmission, see Ben Bernanke and Alan Blinder (1988).

[2] For example, see Stephen King (1986), who examines whether money is a better forecaster of future economic activity than lending.

[3] Benjamin Friedman (1982, 1983, 1986) is an exception in that he does a number of tests aimed at establishing the exogeneity of his credit proxies.

The papers then go on to examine the impulse responses of money, loans, and various measures of economic activity to these indicators. Although different indicators are used (Bernanke and Blinder look at the federal funds rate and its spreads to Treasuries, while Romer and Romer rely on their reading of the minutes of Federal Open Market Committee [FOMC] meetings to identify shifts to tighter policy), many of the qualitative conclusions are similar. When the Fed tightens, the money stock falls almost immediately. Bank loans fall also, but with a significant lag: the decline does not begin to show up for 6–9 months. Output falls with a lag also and indeed seems to move roughly contemporaneously with loans.

However, even if the indicators used in these papers are indeed good proxies for exogenous impulses to monetary policy, an identification problem still remains in terms of deciding whether a distinct lending channel is at work. Just because a fall in output coincides with a fall in loans does not establish that the former was caused by the latter. It is possible that the entire output response to the policy tightening was due to the conventional money channel and that the fall in the quantity of loans simply reflects a decrease in loan demand (due to reduced output) and not a reduction in loan supply.

In this paper, we bring new evidence to bear that allows for a clearer econometric identification of the lending channel of monetary-policy transmission. Rather than just looking at how bank assets and liabilities respond to policy impulses, we also focus on the behavior of an important substitute for bank finance, namely commercial paper. The intuition for why such data can aid in achieving identification is simple. Suppose that monetary policy operated solely through a money channel, and that the fall in bank loans seen when the Fed tightens is due only to an output-induced effect on credit demand. Then one should expect the demand for nonbank sources of credit to decline as well, leading to a reduction in, say, the volume of commercial paper issues. If, on the other hand, Fed tightening reduces the supply of bank credit, one might expect an increase in commercial paper issuance, to the extent that businesses have some ability to substitute between the two sources of finance.

In the next section of the paper, we formalize the links between the stance of monetary policy and the composition of external finance with the aid of a simple model. The model highlights two basic conditions that are required for the existence of a lending channel and illustrates how data on nonbank sources of finance can be used to test each condition. Section II contains the first set of tests, and Section III contains the second set.

In addition to being informative about the transmission of monetary policy, our results also shed new light on a statistical finding that has attracted a great deal of attention in macroeconomics recently, namely, that the spread between prime commercial paper (CP) rates and Treasury bill rates forecasts economic activity surprisingly well. As noted by Friedman and Kenneth Kuttner (1992) and James Stock and Mark Watson (1989), the CP–bills spread has been a very powerful leading indicator.

While the spread's forecasting power appears to be well established, the economics of why it has forecast so well are less clear. One common interpretation is that the spread simply reflects default risk and that this forward-looking property is what makes it a good predictor. In Section IV, we argue against this interpretation. We suggest instead that the CP–bills spread is a proxy for the stance of monetary policy: tight monetary policy leads to an increase in commercial paper issuance, which exerts upward pressure on paper rates.[4] If tight money eventually has an output effect, this effect will have been forecast by the movement in the CP–bills spread.

As we explain in Section IV, the distinction between these two hypotheses is a potentially important one. Our theory leads one to be skeptical about the extent to

[4]Bernanke (1990) also argues that the spread reflects monetary policy; he notes but does not stress the role of commercial paper issuance as we do here.

which the historical correlation between the spread and economic activity can be expected to hold up: as the commercial paper market deepens, the price pressure generated by a given Fed tightening should decline. At the same time, our theory suggests that data on the volume of commercial paper issuance may continue to be a useful leading indicator. We provide some simple forecasting comparisons between interest rate spreads and the financing mix which are consistent with this view.

I. The Model

The model we present below is highly stylized. Its aim is not to be either particularly realistic or rigorous, but rather to illustrate simply the logic behind our empirical tests. The model is very similar to the one in Bernanke and Blinder (1988), with the one key difference being our more explicit focus on firms' capital structure choices. The model highlights the conditions under which all distinctions between bank loans and securities can be ignored, as is done in the standard *IS-LM* model where there are only two financial assets: "money" and "bonds." When these conditions are not satisfied, there are three assets—money, "bank loans," and "securities"—which must be accounted for separately.[5]

We begin by describing the investment and financing behavior of a representative firm. The firm invests an amount I and can choose between two sources of finance in raising the funds for the investment. A fraction α of the financing comes from bank loans at an interest rate r_ℓ, and the remaining fraction $(1-\alpha)$ comes from commercial paper at rate r_p. In order to keep the model simple, we assume that all markets clear by price. However, when we turn to the empirical work, it will be important to recognize that the "true" price of bank loans is imperfectly observable; because of the widespread use of nonprice terms of credit (e.g., covenants, collateral, or the type of quantity rationing discussed in Joseph Stiglitz and Andrew Weiss [1981]), variables such as the prime rate may be a very noisy measure of r_ℓ.

In addition to the direct interest cost, we assume that there is a "relationship" benefit R to bank borrowing that depends on the total amount borrowed from the bank:

$$R = If(\alpha) \tag{1}$$

where $f(\cdot)$ is an increasing concave function.

We do not explicitly model the sources of this relationship benefit here. However, there are several theories that can explain why firms with an option to do so might choose to finance themselves at least partially with higher-rate bank loans. For one, a banking relationship is likely to involve more monitoring and hence a lesser degree of informational asymmetry between borrowers and lenders. In this sense, a banking relationship can function like an extended internal capital market, allowing firms to finance investments even when adverse-selection problems make it difficult to raise funds from the public markets.[6] A banking relationship may also help to reduce the costs of financial distress: the free-rider problems associated with continued funding of a distressed firm are likely to be less severe than in the case when creditors are all dispersed security-holders.[7]

Given our assumptions, the firm's optimal financing mix, which we denote by α^*, is given by

$$\alpha^* = F(r_\ell - r_p) \tag{2}$$

[5]Embedded in this simple version of the model is the (unrealistic) assumption that publicly traded corporate and government securities are perfect substitutes. This is an assumption that we will implicitly be relaxing when we discuss the determinants of the spread between commercial paper and Treasury bills in Section IV.

[6]For an example of theoretical work on the benefits of bank monitoring, see Douglas Diamond (1984, 1991). Takeo Hoshi et al. (1991) find empirical support for the idea that banking relationships can be valuable, because they help to make firms' investment less dependent on their internal cash flow. See also Eugene Fama (1985).

[7]Hoshi et al. (1990) and Stuart Gilson et al. (1990) present evidence consistent with this view.

where $F(\cdot) = f'^{-1}(\cdot)$ is a decreasing function. Equation (2) contains our first important insight. It establishes that any shock (e.g., monetary policy) that disturbs the relative costs of loans and paper will be reflected in shifts in firms' financing mix.[8] Thus, even if the "true" r_ℓ is difficult to observe accurately in practice, we can use data on firms' financing choices to infer something about the state of loan supply.[9]

We now close the model in the simplest possible fashion. Given the firm's choice of capital structure, its net cost of capital (inclusive of the relationship benefit of bank loans), k, is

$$(3) \qquad k = r_p + \alpha^*(r_\ell - r_p) - f(\alpha^*).$$

Investment demand is assumed to depend on both k and aggregate output Y:[10]

$$(4) \qquad I = I^d(Y, k).$$

[8]In our formulation, a change in the relative cost of loans and paper is the *only* thing that alters the mix. Diamond (1991) presents a model in which the mix can be changed due to movements in the relative *demand* schedules for loans and paper. Diamond's model has the feature that R is effectively higher when real interest rates are higher, implying that the mix will *rise* in the wake of a monetary contraction. Therefore, the considerations in Diamond's basic model suggest that our tests will be excessively conservative; it will be harder for us to find the loan-supply effects we are looking for.

[9]There is another reason why looking at quantity financing variables can add value relative to using interest rate spreads as a proxy for $(r_\ell - r_p)$: the observed rate spread may be affected by other factors outside our model, such as changing default probabilities. For example, it may be that in a recession the likelihood of default rises more for the small companies who are bank borrowers than for the large companies who are commercial paper borrowers. If this is the case, the *observed* rate spread will widen even in the absence of any change in the "true" (net of default) costs of loans and paper. This sort of measurement error could be particularly problematic, as it is likely to be correlated with many of the business-cycle variables that we will be studying.

[10]The model is atemporal so investment and the capital stock coincide. Using standard adjustment-costs arguments, we could introduce dynamics and transform the demand for capital into a model for investment.

Aggregate output, in turn, is the sum of investment and other autonomous demand:[11]

$$(5) \qquad Y = I + G.$$

Interest rates on both paper and bank loans are determined by monetary policy. To keep things as simple as possible, we assume that banks are financed completely with demand deposits, hold cash to satisfy reserve requirements, and hold no excess reserves. Thus the quantity of money, M, is a sufficient statistic for the stance of monetary policy.[12] Money affects the interest rate on nonbank paper through the usual *LM*-type relationship:

$$(6) \qquad r_p = H(Y, M).$$

Money can also affect the interest rate on bank loans if banks have nondegenerate portfolio preferences across loans and paper as assets on their balance sheets. We assume that banks wish to hold a fraction of their assets in loans and that this fraction depends on the spread between the rate on loans and the rate on paper. Thus, loan supply L^s is given by

$$(7) \qquad L^s = J(r_\ell - r_p)M$$

where $J(\cdot)$ is an increasing function. Note that since we allow banks to cushion the impact of monetary policy on loans by carrying out portfolio adjustments on the asset side of their balance sheets, there is no real loss of generality in assuming that they are inflexible with respect to the structure of their liabilities.

Equating loan supply to loan demand yields

$$(8) \qquad \alpha^* I = J(r_\ell - r_p)M.$$

[11]For simplicity, we do not model consumption behavior.

[12]This is an obviously unrealistic assumption, and it is made for expository purposes only. Indeed, in the empirical work that follows, we use indicators other than the quantity of money to gauge the stance of monetary policy.

The six equations, (2)–(6) and (8), determine the six endogenous variables Y, I, k, r_ℓ, r_p, and α^* in terms of the exogenous parameters G and M. It is clear from these equations that there are two necessary conditions that must both be satisfied if monetary policy is to impact aggregate demand in part through a distinct lending channel:

(i) Loans and paper must be imperfect substitutes as bank assets. That is, banks must not react to a contraction in size simply by reducing their holdings of paper and leaving loan supply unchanged.
(ii) Loans and paper must also be imperfect substitutes as corporate liabilities. Firms must not be able to offset costlessly a reduction in bank loan supply by issuing more paper.

The next two sections of the paper are devoted to testing each of these necessary conditions in turn. In both cases, the model helps us to see how data on either the volume of nonbank paper or the ratio of bank loans to total financing can help in creating unambiguous tests.

To test condition (i), we investigate how paper volume and the ratio of loans to total financing respond to a monetary-policy impulse. That this provides a sharper test of condition (i) than simply looking at the response of loan volume to a monetary impulse can be seen by computing the following three derivatives:

$$(9)\quad d(\text{loans})/dM = \alpha^* \, dI/dM + I\, d\alpha^*/dM$$

$$(10)\quad d(\text{paper})/dM = (1-\alpha^*)\, dI/dM - I\, d\alpha^*/dM$$

$$(11)\quad d\alpha^*/dM = F'\, d(r_\ell - r_p)/dM.$$

According to equation (11), the financing ratio can vary with money only if condition (i) holds. If condition (i) does not hold, and loans and paper are perfect substitutes as bank assets, then the rates on both will be equalized, and α^* will not be affected by monetary policy. Thus, looking at how α^* moves with M can be informative about whether condition (i) holds.[13] The same cannot be said for loans. Equation (9) makes it clear that even if there is perfect substitutability (so that $d\alpha^*/dM = 0$), loans will be positively correlated with money due to their positive correlation with the level of investment. This is the source of the identification problem noted in the Introduction.

The responsiveness of paper volume to a monetary impulse provides an even stricter test of condition (i) than looking at α^*. As can be seen from equation (10), a monetary tightening has two opposing effects on paper volume. On the one hand, it leads to a decline in investment that reduces the demand for all types of financing. On the other hand, it may lead to a substitution away from loans and toward paper, thereby increasing paper volume. Thus the correlation between paper and money will be negative only if the effect of M on α^* is strong enough to overwhelm the positive effect of money on output and investment.

Our tests of condition (ii) are motivated by the observation that, in a Modigliani-Miller world where firms can costlessly substitute between bank and nonbank sources of finance, their financing choices (as measured by α^*) should be uninformative about investment activity. Thus, we can test condition (ii) by examining the relationship between α^* and various measures of investment.

The rationale for our tests of condition (ii) can be seen more formally by considering the total derivative of investment:

$$(12)\quad dI = I_y dy + I_k dr_p + I_k \alpha^* (dr_\ell - dr_p).$$

In a world where bank loans and paper are

[13] We also check directly to see how an interest-rate-based proxy for $r_\ell - r_p$ (the spread between the prime rate and the CP rate) responds to a monetary policy shock. To the extent that this proxy is an accurate one, we would expect to obtain similar results using both it and the financing ratio. However, as emphasized above, there are various potential measurement-error problems associated with using an observed rate spread as a proxy for the "true" cost differential between loans and paper.

perfect substitutes, the third term on the right-hand side of (12) drops out, and investment depends only on output and the security-market interest rate. However, with imperfect substitutability, the spread between loan and paper costs also affects investment. In testing condition (ii), we estimate investment equations similar in form to (12), making use of changes in α^* to capture changes in the spread between loan and paper costs.[14]

Before proceeding, we ought to comment on the homogeneity assumption implicit in our model, wherein a single representative firm faces a meaningful trade-off between bank loans and commercial paper. More realistically, there will be some heterogeneity: some larger firms may already be financing themselves solely from nonbank sources, and some small firms may be unable ever to access the securities markets. This leaves only some firms in the middle facing a meaningful trade-off.

Such heterogeneity will probably make it *harder* for us to find the evidence we are looking for. To see why, consider the polar case in which all firms are either completely in or completely out of the paper market, so that there are no firms facing a trade-off at the margin. In such a case, monetary policy will indeed have lending-channel effects—it will lead to a credit crunch for bank-dependent firms—but commercial paper issuance will not increase, because the bank-dependent firms cannot move into the paper market.[15] Thus, heterogeneity cannot "explain away" a finding that paper issuance rises in the wake of a monetary tightening. If anything, it makes such a finding all the more striking.

II. Monetary Policy and the Composition of External Finance

In this section, we look at how the volume of bank loans and commercial paper outstanding respond to changes in the stance of monetary policy. Unfortunately, measuring the stance of monetary policy is in itself not a straightforward task. Given that any single measure is imperfect and subject to criticism, we use several different monetary-policy indicators which have been suggested in the literature.

In our first set of tests, we follow the technique of Romer and Romer (1990), who read the minutes of the FOMC and identify six dates since World War II when the Fed appears to have opted for a clear shift to tighter policy. These "Romer dates" are: October 1947, September 1955, December 1968, April 1974, August 1978, and October 1979. Thus, one simple experiment is to compare the behavior of commercial paper and bank loan volume before and after Romer dates.[16] An increase in commercial paper volume subsequent to a Romer date would constitute evidence of a constriction in bank loan supply.

Although intuitively appealing, the Romer-date approach, by focusing on only a handful of extreme episodes, may sacrifice valuable information. To recapture some of that information, we also use data from the federal funds market. Bernanke and Blinder (1992) argue persuasively that the federal funds rate, as well as the spread between federal funds and Treasury bonds, are good indicators of monetary policy. Our second set of tests thus examines the corre-

[14]Again, we also use the prime–CP spread as a proxy for $r_\ell - r_p$ in our investment equations. The same caveats discussed above apply.

[15]This polar version of the model seems to be what Bernanke and Blinder (1988) have in mind.

[16]A couple of questions have been raised concerning the use of Romer dates as proxies for monetary policy. First, it is difficult to make an airtight case for their exogeneity; the Fed may be endogenously responding to changes in activity when it opts to tighten policy. However, we believe that this presents less of a problem for our use of the Romer dates than it might in other applications. For example, the endogeneity issue might be problematic if one were trying to argue that monetary policy has a causal effect on output. If a Romer date endogenously tends to follow temporarily elevated GNP, it might be incorrect then to infer that the tightening causes any subsequent decline in GNP. However, we think it is much less likely that Fed policy responds endogenously to fluctuations in variables such as the financing mix.

A second issue is the Romer dating criterion's potential omission of other significant episodes of policy tightening, such as the 1966 "credit crunch." We shall discuss this issue shortly.

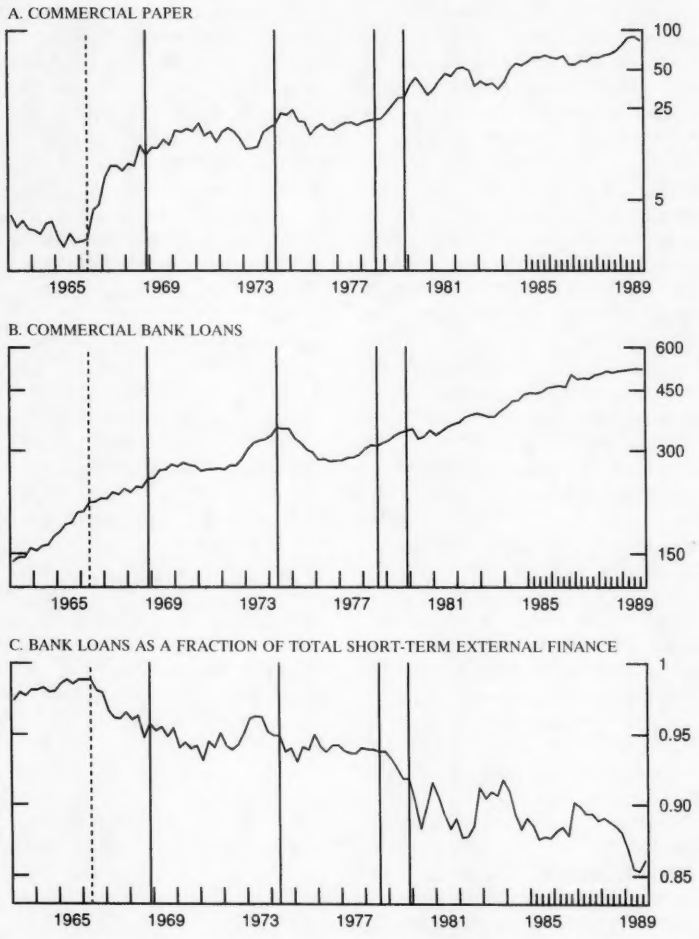


FIGURE 1. THE COMPOSITION OF SHORT-TERM EXTERNAL FINANCE
(FLOW-OF-FUNDS DATA, QUARTERLY, 1963–1989)

Notes: Solid vertical lines mark Romer dates; the dotted line represents the 1966 credit crunch. Nominal values are deflated using the GNP implicit deflator. Commercial paper and bank loan values are in billions of 1982 dollars.

lation between both of these interest rate indicators and the financing variables.

Figure 1 presents the basic financing data that will be analyzed in this section: the log of the real amount of commercial paper issued by nonfinancial domestic corporations ("commercial paper" hereafter); the log of the real amount of commercial bank loans made to businesses ("bank loans" hereafter); and the ratio of the bank loans to the sum of bank loans and commercial paper ("mix" hereafter). These data, along with all the other data used in the paper, are described in the Data Appendix. The

mix variable is an empirical proxy for the quantity α. The solid vertical lines in the chart denote the dates identified by the Romers as the onset of monetary tightening.

The dashed vertical line marks the beginning of the 1966 "credit crunch." Although this episode was not singled out by the Romers, it is widely regarded as one of the most significant periods of tight credit in the post-World War II period. For example, Philip Cagan (1972) notes that monetary growth was flat over the last three quarters of 1966. Moreover, he argues that the decision made by the FOMC on May 10 of that year, to restrict reserve growth until deposit growth slowed, represented a significant tightening of policy. In light of this ambiguity, the formal statistical tests that we will present shortly are done two ways: both with and without the 1966 credit crunch included in our list of tightening episodes.

The figure illustrates the phenomenal growth of the commercial paper market over the last 25 years. As the mix variable shows, commercial paper moved from being a trivial fraction of financing (less than 2 percent) to becoming a large source of financing.

Figure 2 focuses on the movements in commercial paper, bank loans, and the mix in the year before and three years following each of the Romer dates (and the 1966 credit crunch). For comparability across episodes, each series is normalized to equal zero as of the date of the monetary tightening. For reference, the straight line in each graph shows the sample average trend in each series. The average across the Romer episodes (not including the 1966 episode) is given by the bold line in each panel.

Panel A of Figure 2 confirms that across all of the episodes, paper grows at or above trend over the first year following the focal dates. The 1966 experience is particularly noteworthy, as paper volume soared over the last half of 1966 and the beginning of 1967. Interestingly, many analysts studying this period explain the boom in commercial paper issuance using exactly the sort of logic discussed above: they argue that the paper market took off when high-quality borrowers seeking loans were turned away by their bankers. For instance, Timothy Rowe (1986, p. 119) reports that during the 1966 credit crunch, "many potential borrowers who formerly relied on bank short-term credit were forced to turn to the commercial paper market. Consequently, the annual growth rate of commercial paper outstanding rose from 7.8 percent in 1965 to 46.6 percent in 1966."

Returning to the other episodes, the rapid growth in paper tends to abate over the second and third years following the focal date. Panel B of Figure 2 shows that there is relatively little action in bank loans immediately following the Romer dates: as documented by Romer and Romer (1990), bank loans grow at about trend rates for the first few quarters after a monetary contraction. Subsequently, loan growth begins to slow. Bernanke and Blinder (1992) argue that this pattern should be expected since loan commitments and other implicit arrangements prevent banks from quickly shifting their portfolios in response to a shock. By two full years after each of the Romer dates, loan volume is typically below trend.[17]

The behavior of the mix is as would be expected from the behavior of its individual components. This can be seen in panel C of Figure 2. The mix declines noticeably within the first year after a Romer date, and by the third year it is still below the trend line. Notice again that the 1966 episode stands out as a period of dramatic shifting away from bank financing and toward commercial paper financing.

In panel A of Table 1, we attempt to quantify the statistical significance of these effects. We begin by creating two dummy variables. The first, labeled the "Romer dates" dummy, equals 1 on a Romer date, and 0 otherwise. The second, labeled the "Romer dates plus 1966" dummy, also equals 1 at the onset of the 1966 credit crunch. We then do Granger causality tests to see if movements in either of these variables help forecast movements in firms' financing choices. We do the tests in two

[17]The average decline in bank loans is largely attributable to the 1974 episode, as can be seen from Figure 2.

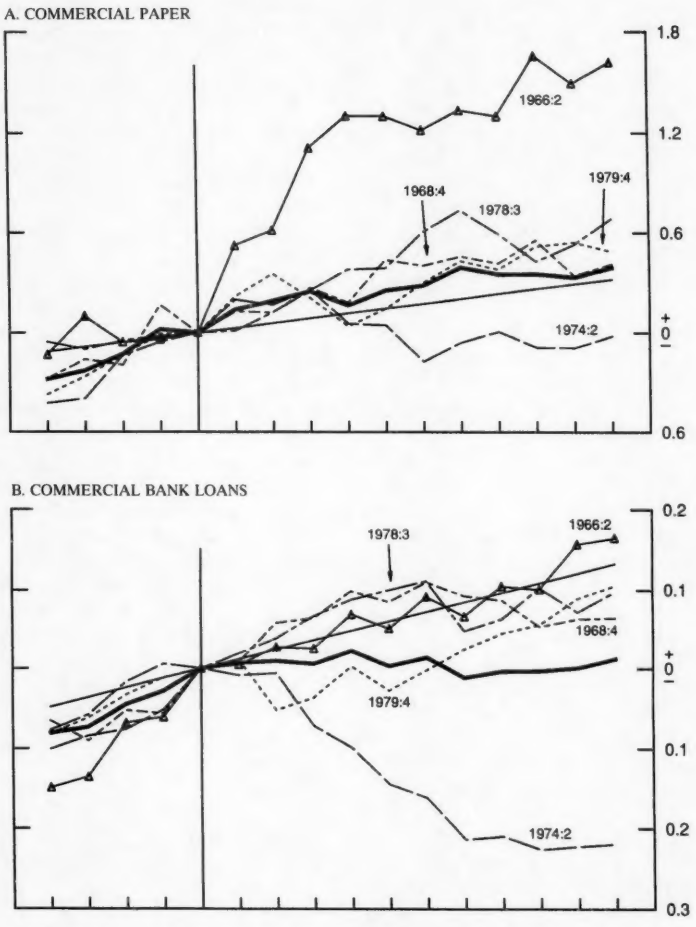


FIGURE 2. CHANGES IN SHORT-TERM EXTERNAL FINANCE AROUND ROMER DATES

Notes: All series are shown as log deviations from their values as of the Romer dates. Ticks along the *x*-axis mark quarters. Averages across Romer dates, excluding the 1966 episode, are shown by the bold lines in each panel. Average growth over 1963–1989 is shown as the straight line.

basic ways. In the "bivariate" version of the tests, we regress the change in the financing variable (e.g., the mix) on eight lags of itself and on eight lags of the dummy variable. In the "multivariate" version, we add eight lags of GNP growth to the equation, in an effort to control for cyclical factors other than monetary policy which might conceivably affect the financing variables. Thus for each financing variable (mix, paper, and loans) we conduct four separate tests.

The results in panel A of Table 1 indicate that many of the patterns displayed in Figure 2 are indeed statistically significant, in

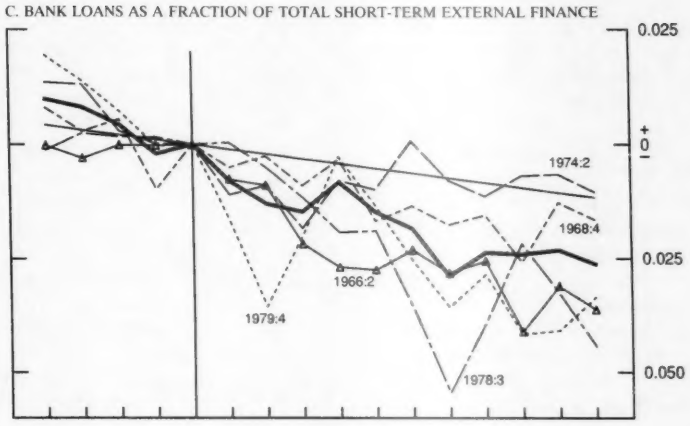


FIGURE 2. (*Continued*)

TABLE 1—CAUSALITY TESTS BETWEEN FINANCING VARIABLES AND INDICATORS OF THE STANCE OF MONETARY POLICY

Indicator	Mix		Log real paper		Log real loans		Prime–CP spread	
	Exclusion	Summation	Exclusion	Summation	Exclusion	Summation	Exclusion	Summation
A. *Romer Date Indicators:*								
Romer dates (bivariate)	0.09	−2.10	0.84	0.61	0.17	−2.16	0.0002	2.17
Romer dates plus 1966 (bivariate)	0.08	−2.55	0.13	1.99	0.07	−1.99	0.002	2.27
Romer dates (multivariate)	0.02	−2.91	0.57	1.63	0.45	−1.02	0.0009	2.35
Romer dates plus 1966 (multivariate)	0.05	−3.02	0.07	2.96	0.12	−1.18	0.02	1.98
B. *Interest Rate Indicators:*								
Funds rate (bivariate)	0.003	−3.06	0.06	2.23	0.24	−1.46	0.001	4.10
Spread: 10-year government bond–funds (bivariate)	0.03	1.92	0.14	−1.10	0.55	1.61	0.003	−1.48
Funds rate (multivariate)	0.007	−1.92	0.20	1.35	0.58	−0.15	0.002	2.44
Spread: 10-year government bond–funds (multivariate)	0.04	2.36	0.10	−2.73	0.99	−0.48	0.028	−1.10

Notes: In each case, the variable at the top of the column is regressed against eight lags of itself and eight lags of the variable in the row. In the multivariate cases, eight lags of GNP growth are added to the regression. The following variables have been differenced so that they enter the regressions in stationary form: mix, log real paper, log real loans, prime–CP spread, and funds rate. Entries in the "exclusion" columns report the marginal significance levels on tests that variables shown in rows do not help forecast the variable at the top of the column. Entries in the "summation" columns report the t statistic for the test that the sum of the coefficients on the eight lags of the variable shown in the row is zero.

spite of the small number of observations. The sum of the coefficients on the dummy variable is significantly negative in all four of the mix regressions. Thus, even without including the 1966 episode, we find that the mix falls significantly after a Romer date.

The results for the individual components of the mix (i.e., loans and commercial paper) are somewhat less uniform, although they too always go in the predicted direction. The sum of the coefficients is positive in all four of the regressions involving commercial paper, but the results are statistically significant only when the 1966 credit crunch is included. In the loan regressions, the coefficients are always negative, but they

are only significant in the bivariate specifications. Thus, while loans appear to decline in the eight quarters following a monetary contraction, this decline is not significantly greater than would be expected given the weakening of GNP that occurs in this period.

Finally, we also tested to see whether the spread between the prime rate and the commercial paper rate widens after the Fed tightens. If the sorts of measurement-error problems discussed above are not too severe, one would expect the behavior of this spread to be closely related to that of the mix. Interestingly, the last column of Table 1 suggests that this is indeed the case: in all four specifications, we observe a statistically significant increase in the prime–CP spread subsequent to a monetary contraction.

Next, we turn to the two interest-rate-based measures of monetary policy proposed by Bernanke and Blinder (1992): the federal funds rate (Bernanke and Blinder's preferred proxy for monetary policy) and the difference between the 10-year constant-maturity government bond rate and the funds rate. We use exactly the same format as above, simply replacing the two dummy-variable series in the regression tests with these two interest-rate series.[18] As before, we conduct both bivariate and multivariate tests, so that we again have four separate tests for each financing variable.

The results of these tests are reported in panel B of Table 1, and they parallel those in panel A quite closely. Most notably, the sum of the coefficients on the monetary-policy indicator is again statistically significant (and of the predicted sign) in all four of the specifications involving the mix variable: the mix falls significantly when either the funds rate rises or when the (bond–funds) spread shrinks.

Commercial paper issuance goes up when either of the interest rate indicators points to tighter monetary policy, although here (as in panel A) the statistical significance depends somewhat on the specification employed. The results for loans are the weakest: they are of marginal significance in the bivariate tests and are completely insignificant in the multivariate tests. Thus, as in panel A, the evidence suggests that the decline in loans in the first eight quarters after a monetary tightening is not much greater than would be expected given the weakening of GNP that occurs in this period.

Finally, as in panel A, the results in panel B also suggest that the (prime–CP) spread widens in the wake of a monetary contraction.

Thus far, we have looked only at the commercial paper market for evidence on the behavior of nonbank sources of finance. While commercial paper may be the closest substitute for bank loans, our narrow focus on this market nonetheless raises a question: what if the increases in commercial paper volume that we document do not represent substitutions away from bank loans, but rather substitutions away from other nonbank sources of finance? If this were the case, it would be inappropriate for us to interpret our results as evidence of a reduction in bank loan supply.

Ideally, we would tackle this issue by analyzing the movements in all major nonbank sources of finance. Commercial paper and trade credit stand out as the leading short-term substitutes for bank loans. Unfortunately, however, the standard aggregate data on gross trade credit is not appropriate for our purposes, since gross trade credit does not represent a *net* source of financing to the corporate sector.[19]

Data availability is less of a problem with respect to long-term financing instruments.

[18]We actually used the *change* in the federal funds rate in the regressions, as the funds rate itself is nonstationary. Thus, all the tests in panels A and B of Table 1 involve differencing every nonstationary variable. However, we have also rerun all the tests with an alternative specification in which the nonstationary variables enter in level form, but in which a time trend is added to the right-hand side of the equation. This alternative specification produces results very similar to those we report here.

[19]The lending view of monetary-policy transmission has implications for the pattern of *interfirm* trade credit flows, rather than for the aggregate quantity of gross trade credit. For example, if loan supply is constricted, we might expect to see larger firms with access to public debt markets extending more trade credit to smaller firms, an argument made by Allan Meltzer (1960). Pursuing this issue requires firm-level data and is beyond the scope of this paper.

We are not aware of any good theoretical reasons to believe that a substitution from bonds (or other long-term sources of financing) to commercial paper *should* follow a monetary contraction.[20] Nevertheless, it seems worth checking out this possibility empirically. To do so, we looked at new issues of high-grade industrial bonds and performed tests similar to those done above for the other financing variables. The results do not support the notion that a monetary contraction leads to a swap from bonds to paper. For example, we find a *positive* (albeit statistically insignificant) correlation between the federal funds rate and bond issuance. In other words, bonds typically move in the same direction as paper after the Fed tightens. This strengthens the case for believing that the movements seen in commercial paper volume are due to loan-supply effects rather than to relative shifts in the demand for alternative sources of nonbank finance.

III. The Composition of External Finance and Real Activity

We now turn to an investigation of whether condition (ii)—imperfect substitutability by corporations between loans and paper—is satisfied. Our tests are motivated by the observation that in a Modigliani-Miller world where businesses can costlessly substitute between bank and nonbank sources of finance, their financing choices should be uninformative about investment activity. In contrast, our model suggests that data on financing choices should have some predictive power for investment.

Our basic approach is to ask whether the financing mix adds significant explanatory power to investment equations that already include the usual interest rate variables. If we did not control for interest rates, our results would not be able to verify the existence of an independent lending channel of transmission. For example, an unconditional correlation between the mix and real activity could arise if monetary policy moved the mix around [i.e., condition (i) was satisfied] but influenced the real economy solely through its effect on security-market interest rates [i.e., condition (ii) failed to hold].[21]

Unfortunately, the empirical success of most investment and inventory models is open to question. Furthermore, simultaneity and feedback relationships tend to make the results from reduced-form models such as VAR's difficult to interpret. We proceed by examining several "off-the-shelf" structural models for inventories and investment. The advantage of this approach is that it makes the interpretation of the results much more straightforward. The cost is that one must be more concrete in specifying how the economy operates. To avoid being tied to one point of view on this question, we use a host of models.

In the case of fixed investment, there are several well-known empirical models. We examine three: the "accelerator," the "neoclassical," and the securities-value (or "Q") models. Our specifications follow directly from the work of Peter Clark (1979), who conducts a detailed comparison of these models as well as a couple of others. The first two models, the neoclassical and the accelerator, can be motivated from a partial-adjustment mechanism. If we denote the desired capital stock by K^*, the accelerator model is derived by assuming that this desired level is proportional to the current level of output, Y, and that some costs of adjustment impede firms from instantaneously equating their actual capital stock with their desired stock. Given the slow

[20] For example, a simple expectational model of the term structure provides no theoretical guidance as to how changes in monetary policy and the accompanying interest rate changes should affect firms' choice of debt maturity.

[21] This is not to imply that unconditional correlations between the mix and real activity are uninteresting. Indeed, we examine such correlations in the next section and compare them to the well-documented correlations between the CP–bills spread and real activity. Our only point here is that unconditional correlations are inappropriate for drawing a clear distinction between money and credit channels.

adjustment, net investment I^N will be given by

$$(13) \qquad I_t^N = \mu + \sum_{s=0}^{N} \beta_s \Delta K_{t-s}^* = \mu + \sum_{s=0}^{N} \beta_s \Delta Y_{t-s}$$

where Δ represents the first-difference operator and N is the length of the distributed lag. To model gross investment, I, we assume that replacement demand is proportional to the lagged capital stock. Furthermore, we follow Clark (1979) in normalizing all variables by potential output, YP, to correct for heteroscedasticity, and in assuming that the stochastic disturbance in the equation exhibits first-order serial correlation.[22] Therefore the specification of the Clark accelerator equation is

$$(14) \quad \left(\frac{I}{\mathrm{YP}}\right)_t = \frac{\mu}{\mathrm{YP}_t} + \sum_{s=0}^{N} \beta_s \left(\frac{\Delta Y_{t-s}}{\mathrm{YP}_{t-s}}\right) + \frac{\delta K_{t-1}}{\mathrm{YP}_t} + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t.$$

The Jorgenson-style neoclassical model that we consider is obtained by postulating that the desired capital stock is the following linear function of output:

$$(15) \qquad K^* = \gamma \frac{pY}{C}$$

where p is the price of output and C is the rental price of capital services. This formula can be derived by assuming that output is produced competitively and that the production function is Cobb-Douglas. In this case, the parameter γ is the share of capital in output. On theoretical grounds, many prefer the neoclassical model over the accelerator model since it introduces a dependency between investment and the cost of capital. Given the earlier assumptions regarding replacement investment, heteroscedasticity, and serial correlation, the specification Clark (1979) analyzes is:

$$(16) \quad \left(\frac{I}{\mathrm{YP}}\right)_t = \frac{\mu}{\mathrm{YP}_t} + \sum_{s=0}^{N} \beta_s \left[\frac{\Delta\left(\frac{pY}{C}\right)_{t-s}}{\mathrm{YP}_{t-s}}\right] + \frac{\delta K_{t-1}}{\mathrm{YP}_t} + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t.$$

We also experimented with Clark's version of Bischoff's modified neoclassical model, which allows for different coefficients on output and the cost of capital. The results were very similar to the results from the simple neoclassical specification and thus are omitted.

Our final model, the Q model, is based on the assumption that if the market value of a firm exceeds the replacement cost of its assets, then it should seek to expand. Our specification, again taken from Clark (1979), is

$$(17) \qquad \frac{I_t}{K_{t-1}} = \mu + \sum_{s=0}^{N} \beta_s Q_{t-s} + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t$$

where Q is the market value of the firm (corrected for the effect of taxes) divided by the replacement cost of the firm. Together these three models roughly encompass most of the structural equations that have been proposed for studying investment.

There is even less consensus on how to handle inventories. We follow Louis Maccini and Robert Rossana (1981) and embed the role of interest rates in a partial-adjustment model. As Blinder and Maccini (1991)

[22] Without scaling, the residuals from the equation seem to be quite heteroscedastic. In a large sample, the scaling should only help to correct the heteroscedasticity and not otherwise affect the equation. We find that the scaling not only helps to correct the heteroscedasticity, but also makes the impact of the mix more pronounced in our augmented equations.

note, although this strategy has been relatively unsuccessful in finding interest rate effects, it is virtually the only tractable alternative and therefore has been widely used. Letting H be the actual stock of inventories and H^* be the desired stock of inventories, Maccini and Rossana begin with

$$(18)\quad \Delta \log(H_t) = \lambda[\log(H_t^*) - \log(H_{t-1})]$$

$$(19)\quad \log(H_t^*) = \mu + \beta \log(S_t^e) + \gamma r_t^e$$

where S^e and r^e are expected sales and carrying costs, the factors that are typically assumed to be the key determinants of desired inventories.[23] Maccini and Rossana follow the standard practice of using distributed lags to proxy for the expected value of sales and carrying costs. Thus, after substituting to eliminate the desired target and allowing for serial correlation, the equation they consider is

$$(20)\quad \Delta \log(H_t) = \mu + \sum_{s=1}^{N_1} \beta_s \log(S_{t-s}) + \sum_{s=1}^{N_2} \gamma_s r_{t-s} + \delta \log(H_{t-1}) + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t$$

where S and r are actual sales and interest costs. We use two variants of this equation. The first, which we label the "accelerator" model, contains only the terms pertaining to expected sales. The second, labeled "neoclassical" in what follows, includes both the sales and interest-expense proxies.

Before the equations can be estimated several other details must be finalized. First, we need to identify the lag lengths that will be used in the estimation. For the investment equations we follow Clark (1979) by using quarterly data and allowing for 20 lags on the output and cost-of-capital measures, but restricting the lag coefficients to lie on a sixth-order polynomial. In the inventory equations we follow Maccini and Rossana (1981) by allowing for 24 lags of sales and 20 lags for the inventory carrying costs and restricting these coefficients to lie on a third-order polynomial. (Limited experimentation suggested that the results were robust with respect to both of these choices.)

Second, we must specify which inventory and investment series will be used. For the investment series, we follow the standard convention in the literature and distinguish between investment in equipment and investment in nonresidential structures. For inventories, we follow the distinction made in the national income and product accounts and isolate inventories held by manufacturers, wholesalers, and retailers of nondurable goods and inventories of manufacturers and wholesalers of durable goods.[24] For convenience, we refer to these as nondurables and durables inventories, respectively.

Finally, to test whether the mix has any additional explanatory power, we added the mix to each equation with a specification symmetric to that for interest rates. For the investment equations this amounts to adding 20 lags of the mix with the coefficients constrained to lie on a sixth-order polynomial, while for inventories this strategy suggests allowing for 20 lags of the mix with the coefficients constrained to lie on a third-order polynomial. (We obtained fairly similar results when we instead added four lags of the mix to the equations and did not restrict the coefficients.) The equations are estimated over the 1964–1988 period.[25]

[23] Maccini and Rossana (1981) also experiment with allowing inventories to depend on other factors such as materials costs and wages. Those extensions had little effect on the importance of interest rate effects, so we suppress them.

[24] Thus we are ignoring retail durable-goods inventories. This category is mostly automobiles and is relatively small compared to the other two categories we consider.

[25] The sample period was chosen to correspond to the availability of the data for Q, which was available only through 1988.

TABLE 2—TESTS FOR EXPLANATORY POWER OF MIX AND PRIME–CP SPREAD IN STRUCTURAL INVENTORY AND INVESTMENT EQUATIONS

Category	Mix			Prime–CP spread		
	Accelerator	Neoclassical	*Q*	Accelerator	Neoclassical	*Q*
Nondurable inventories	0.173	0.009	—	0.022	0.001	—
Durable inventories	0.002	0.049	—	0.500	0.302	—
Producer's durable equipment	10^{-6}	0.003	0.026	10^{-7}	0.003	0.027
Nonresidential structures	0.460	0.391	0.175	0.537	0.011	0.037

Notes: The number in each cell is the significance level in an *F* test that tests whether a distributed lag of either the mix or the prime–CP spread enters the inventory and investment equations described in the text. "Nondurable inventories" are inventories held by manufacturers, wholesalers, and retailers of nondurable goods; "durable inventories" are inventories held by manufacturers and wholesalers of durable goods.

In Table 2 we report the significance levels for the *F* tests of whether all the coefficients on the mix are jointly equal to zero.[26] The results suggest that the mix is an important determinant of inventories; the mix shows up very strongly for the durables inventories and moderately strongly for the nondurables inventories. The shapes of the lag distributions imply that the mix has the expected effect, at least over the first couple of years: a declining loan share is associated with falling inventories. Given that the long-run output effects of monetary policy would be predicted to be small, it is also not surprising to find that the long-run impact of a shift in the mix is insignificantly different from zero.

These results—that inventories can be influenced by financial factors—are particularly interesting in light of two stylized facts. First, interest rate effects have been hard to find in standard inventory models. Second, there is a very tight positive correlation between bank loans and inventories (see e.g., Board of Governors of the Federal Reserve System, 1980 chart 1). By itself, of course, this correlation says nothing about any causality running from loans to inventories; it could just reflect fluctuations in inventory demand that are passively accommodated by bank loan supply. However, our results suggest that inventories could indeed be affected by the availability of bank credit. Thus financing costs may be important for inventories, but these costs may not be adequately captured by security-market interest rates alone.

The results for investment are less uniform. For equipment investment, the mix again appears to be very important, regardless of the baseline model chosen. In contrast, the mix appears to be much less important for investment in nonresidential structures.[27] Nevertheless, in all cases, the impact of declines in intermediated financing is as expected: there is a negative short-run effect on investment, and this effect dies off at longer horizons.

Table 2 also contains a second set of tests, in which the prime–CP spread is substituted for the mix everywhere. As in the previous section of the paper, we find some evidence that the two variables contain similar information. Like the mix, the prime–CP spread has significant explanatory power both for nondurable inventories and for

[26]The tests are carried out in terms of the level of the mix. The arguments in Jeffrey Wooldridge (1990) suggest that standard asymptotics apply, provided that the mix, conditional on the other right-hand-side variables in the equations is stationary. This condition is easily satisfied in all of the specifications aside from the *Q* equations, where it is marginally satisfied. Nonetheless, to ensure that our small-sample results are not due to any leftover time trends in the investment series that are correlated with the trend in the mix, we reran all the tests, adding a time trend to the right-hand side of each equation. We obtained qualitatively similar results with this procedure.

[27]This finding accords with the common result in the literature that it is notoriously difficult to find important financing effects for structures (see e.g., Robert J. Gordon and John M. Veitch, 1986).

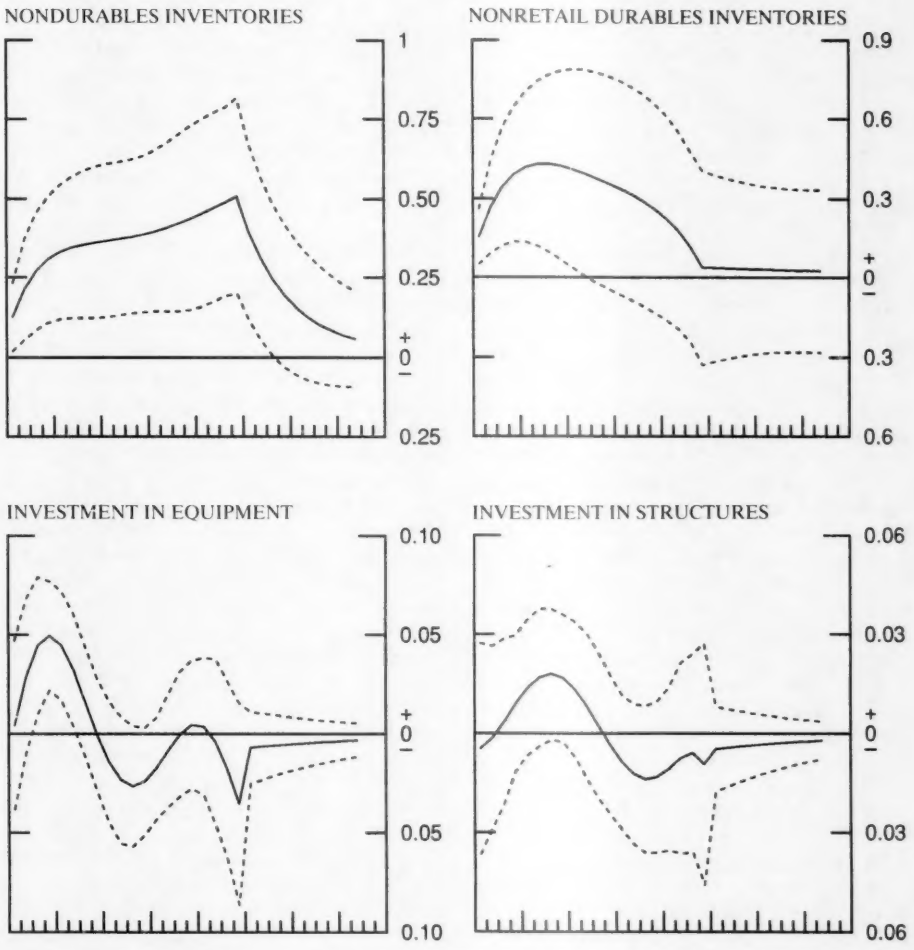


FIGURE 3. IMPULSE RESPONSES AND APPROXIMATE 95-PERCENT CONFIDENCE INTERVALS FOR SHOCKS TO THE MIX IN NEOCLASSICAL MODELS

Notes: Ticks along the *x*-axis mark quarters after the shock to the financing mix. Dotted lines show plus and minus two standard errors around the means, shown as solid lines. The means and confidence intervals were calculated from a Monte Carlo simulation using 500 replications.

equipment investment. Unlike the mix, however, it adds little to the equations for durable inventories. Also unlike the mix, it is significant in two of the three structures specifications.

While Table 2 demonstrates the statistical significance of our loan-supply variables, it does not provide any guidance on their quantitative importance. We now attempt a rough calibration of these magnitudes. To begin, we compute the responses of investment and inventories to shocks in the mix using the neoclassical specifications [equations (16) and (20) augmented to include

lags of the mix]. We also compute confidence intervals using Monte Carlo simulations.[28] The mean responses for a one-unit increase in the mix, along with standard-error bands are shown in Figure 3. As is typical, the range associated with plus and minus two standard errors around the path is fairly wide in all four cases, but at least in the short run, there appear to be some meaningful patterns in the data.

Next, we calculate the impact of a one-time, 0.036 decline in the mix. This is a large decline, comparable to that seen in the two quarters after the Fed's October 1979 tightening. One year after such a shock, inventory investment would be reduced by about $36 billion (in 1982 dollars), equipment investment would be reduced by about $7.5 billion, and structures investment would be reduced by about $1.5 billion. Together this suggests a decline of roughly $45 billion, which is just over 1 percent of GNP. While the standard errors on these estimates are large, the overall magnitudes seem reasonable, and the sectoral incidence is plausible: inventories are most strongly affected, while equipment investment is slightly affected, and investment in structures is hardly affected.

IV. Financing Variables as Predictors of Economic Activity

In the previous section, we showed that financing variables such as our mix variable could help to predict certain components of economic activity, even after controlling for interest rates and other factors. These controls were necessary to address the specific hypothesis we were testing: whether condition (ii) was satisfied and there was thus evidence of a distinct lending channel of monetary transmission.

In this section, we shift our focus away from the issue of trying to differentiate carefully between the money and lending channels and address a less structured set of forecasting questions. In particular, we consider how our mix variable performs as a univariate "leading indicator" and contrast it with another variable that has received a great deal of attention in this regard: the spread between the rate on prime commercial paper and Treasury bills.

Our results thus far suggest a couple of reasons why the mix might be a good univariate leading indicator. First, the evidence in Section III tells us that the mix has some forecasting power even when many other variables are included in the model. Second, the evidence in Section II implies that the mix might also do well in univariate forecasting even if we had found no role for it in the multivariate setting. This is because even if condition (ii) fails to hold, and monetary policy thus works solely through a money channel, we know that the mix is correlated with changes in monetary policy.

The insight that the mix may be a good proxy for the stance of monetary policy, no matter what the eventual channels of policy transmission, is also helpful in understanding the success of the CP–bills spread as a leading indicator. Several explanations for the spread's success have been offered. The first is that the spread reflects expected future default risk and that this forward-looking property is what makes it a good predictor. Friedman and Kuttner (1993), as well as Bernanke (1990), argue that this is at best a partial explanation. For example, Bernanke points out that it is hard to reconcile the 300-basis-point swings seen in the spread with changes in default expectations, given that defaults on prime commercial paper are extremely rare. Furthermore, both papers note that the CP–bills spread is not closely correlated with other more natural measures of default risk, such as the spread between BAA corporate bonds and Treasury bonds, or the spread between low-grade and high-grade commercial paper.

A second possible explanation, put forward by Friedman and Kuttner (1993), focuses on the changes in companies' demand for funds that occur around cyclical turning points. To the extent that the end of an expansion is characterized by a buildup in inventories, there will be an increased demand for such short-term sources of funds

[28] The results are based on 500 replications, assuming that the errors in each equation are normally distributed.

TABLE 3—CAUSALITY TESTS FOR MIX AND SIX-MONTH CP–BILLS SPREAD

	1964–1989		1964–1979:3		1979:4–1989	
Dependent variable	Spread	Mix	Spread	Mix	Spread	Mix
Capacity utilization	10^{-8}	0.048	10^{-8}	0.426	0.110	0.007
Log employment	10^{-8}	0.003	10^{-8}	0.063	0.005	0.078
Log housing starts	0.015	0.105	0.059	0.660	0.229	0.054
Log personal consumption	0.00002	0.314	0.001	0.127	0.049	0.460
Log personal income	0.005	0.572	0.021	0.438	0.423	0.359
Unemployment	10^{-8}	0.017	10^{-8}	0.266	0.012	0.006
Industrial production	10^{-8}	0.007	10^{-8}	0.221	0.017	0.004
Log durable goods orders	10^{-8}	0.067	10^{-6}	0.278	0.036	0.161
Log retail sales	0.0002	0.645	0.003	0.635	0.049	0.529

Notes: Entry in each cell is the significance level from a test that four lags of the mix or spread add explanatory power to a regression that already includes four lags of the dependent variable. All nonstationary variables have been differenced.

as commercial paper. If commercial paper and Treasury bills are imperfect substitutes, this demand-side effect should widen the spread between the two. Friedman and Kuttner (1993) find empirical support for this proposition.

A third explanation for the predictive prowess of the spread—discussed by both Friedman and Kuttner (1993) and Bernanke (1990)—is that the spread contains information about the stance of monetary policy. Our results in Section II support this hypothesis. Specifically, our results point to one clear reason why the CP–bills spread may be a good proxy for the stance of monetary policy: if tight money leads to an increase in commercial paper issuance, and if the commercial paper market is less than perfectly "deep," then the end result will be an increase in paper rates relative to Treasury rates. In other words, the spread may be a good proxy for the stance of monetary policy for exactly the same reasons that the mix is.

However, this line of reasoning raises the following question: should one expect the spread to continue to be a good forecaster in the future? Bernanke (1990) also provides evidence on this point. He shows that the forecasting ability of the spread has deteriorated noticeably over the 1980's. According to the story above, this might be expected if the spread's variation is driven by fluctuations in paper issuance. Over time, as markets become deeper and more liquid, this "price-pressure" effect may be reduced. Even if monetary policy continues to have a significant impact on paper volume, the impact on paper rates may diminish. This would weaken the forecasting power of the spread, while leaving the forecasting power of the mix unaffected.[29]

To investigate this hypothesis, we compare the forecasting properties of the CP–bills spread and our mix variable for a host of indicators of economic activity over several sample periods. Specifically, we follow Bernanke and Blinder (1992) and study the following activity measures: industrial production, capacity utilization, employment, the unemployment rate, housing starts, personal income, retail sales, consumption, and durable-goods orders. We consider three sample periods: 1964–1989, 1964–1979:3 and 1979:4–1989. The sample break point was suggested by Bernanke's (1990) work on the changing forecasting ability of the CP–bills spread.[30]

In each regression, we run the activity variable on four lags of itself and four lags of either the mix or the spread.[31] These results are shown in Table 3. The results for

[29]Bernanke (1990) describes other reasons why market developments might lead to a reduction in the forecasting power of the spread.

[30]By starting in 1964 we capture the 1966 episode without using too much data from the period before the first major expansion in the paper market.

[31]All nonstationary variables have been differenced in these regressions.

the CP–bills spread confirm Bernanke's (1990) findings that the information content of the spread has fallen off noticeably in the 1980's; indeed, the spread's explanatory power declines in the later subsample for every activity variable we study. In most cases, the spread still seems to have some predictive power, but overall the results are much less impressive.

In contrast, the forecasting performance of the mix improves over the later subsample for seven of the nine activity variables. Moreover, in the later sample period, the mix actually outperforms the spread for five of the nine activity measures in these forecasting contests. This is in sharp contrast to the earlier sample period, where the spread uniformly dominated the mix. On balance, these results do suggest that the information content in the mix has held up better than that of the spread as the commercial paper market has deepened. Still, even over the 1980's, it does not appear that the mix is an unambiguously better leading indicator than the spread.

V. Conclusions

There are two necessary conditions that must be satisfied in order for monetary policy to affect the economy through a lending channel. First, banks must view loans and securities as imperfect substitutes on the asset side of their balance sheets, so that monetary tightening does indeed reduce the supply of bank loans. Second, loans and nonbank sources of finance must also be imperfect substitutes for firms on the liability side of their balance sheets, so that reduced loan supply has real effects. The data we have examined suggest that both conditions are satisfied: shifts in monetary policy seem to alter the mix of loans and commercial paper, and the induced shifts in this mix seem to affect investment (even controlling for interest rates).

In this paper, we have looked only at aggregate data. However, the lending view of monetary-policy transmission also has a number of important cross-sectional implications. To take just one example, the lending view suggests that the inventory and investment declines that follow a monetary contraction should be disproportionately concentrated among "bank-dependent" firms (i.e., firms with scarce cash reserves and without ready access to public bond and commercial paper markets). In current work (Kashyap et al., 1992) firm-level data are being used to investigate this hypothesis.

Data Appendix

The data used in the paper can be divided into four categories. Except where otherwise noted, all data come from the Federal Reserve data base.

Financing Measures.—The bank loan and commercial paper series used in our investigation are taken from the Federal Reserve Board's (1980) flow-of-funds publication. The bank loan series is the sum of two series. The first series is "Bank Loans, NEC," line 25, in the "Nonfarm Non-Corporate Business" sector. The second series is "Bank Loans, NEC," line 34, in the "Nonfinancial Corporate Business" sector, excluding farms. The commercial paper series is taken from line 36 of the "Nonfinancial Corporate Business" sector, excluding farms. Finally, the bond issuance series was constructed by splicing a Moody's series and one tracked by the Federal Reserve. The nominal quantity data were adjusted using the 1982-based GNP deflator.

Interest Rates.—We use a variety of interest rates in this study: the federal funds rate, the six-month Treasury bill rate, six-month high-quality commercial paper rate, the prime rate, and the 10-year constant-maturity government security. These are all standard series.

Inventory and Investment Data.—The variables needed to run the inventory equations come from the MPS quarterly-model data base and the National Income and Product Accounts. The MPS data are described at length in Flint Brayton and Eileen Mauskopf (1985) and are available upon request. The data on Q and investment were kindly provided by Glenn D. Rudebusch and are described in Stephen D. Oliner et al. (1992).

Business-Cycle Indicators.—The data for the indicators of economic activity used in Section IV are all available from a number of sources. We used the Federal Reserve Board data base; Bernanke and Blinder (1992) use the same data (obtained from Data Resources, Inc.) and provide the DRI mnemonics.

REFERENCES

Bernanke, Ben S., "On the Predictive Power of Interest Rates and Interest Rate Spreads," *New England Economic Review*, November-December 1990, 51–68.

Bernanke, Ben S. and Blinder, Alan S., "Credit, Money, and Aggregate Demand," *American Economic Review*, May 1988 (*Papers*

and Proceedings), *78*, 435–9.

_____ **and** _____, "The Federal Funds Rate and the Channels of Monetary Transmission," *American Economic Review*, September 1992, *82*, 901–21.

Blinder, Alan and Maccini, Louis J., "Taking Stock: A Critical Assessment of Recent Research on Inventories," *Journal of Economic Perspectives*, Winter 1991, *5*, 73–96.

Brainard, William, "Financial Intermediaries and a Theory of Monetary Control," *Yale Economic Essays*, Fall 1964, *4*, 431–82.

Brayton, Flint and Mauskopf, Eileen, "The Federal Reserve Board MPS Quarterly Econometric Model of the U.S. Economy," *Economic Modelling*, July 1985, *2*, 170–292.

Cagan, Phillip, "Monetary Policy," in Phillip Cagan, Martin Estey, William Fellner, Charles McClure, Jr., and Thomas Gale Moore, eds., *Economic Policy and Inflation in the Sixties*, Washington, DC: American Enterprise Institute for Public Policy Research, 1972, pp. 89–153.

Clark, Peter, "Investment in the 1970s: Theory, Performance, and Prediction," *Brookings Papers on Economic Activity*, 1979, (1), 73–113.

Diamond, Douglas, "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, July 1984, *51*, 393–414.

_____, "Monitoring and Reputation: The Choice between Bank Loans and Directly Placed Debt," *Journal of Political Economy*, August 1991, *99*, 689–721.

Fama, Eugene F., "What's Different About Banks?" *Journal of Monetary Economics*, January 1985, *15*, 29–39.

Friedman, Benjamin M., "Debt and Economic Activity in the United States," in Benjamin M. Friedman, ed., *The Changing Roles of Debt and Equity in Financing U.S. Capital Formation*, Chicago: University of Chicago Press, 1982, pp. 91–110.

_____, "The Roles of Money and Credit in Macroeconomic Analysis," in James Tobin, ed., *Macroeconomics, Prices, and Quantities: Essays in Memory of Arthur M. Okun*, Washington, DC: Brookings Institution, 1983, pp. 161–89.

_____, "Money, Credit, and Interest Rates in the Business Cycle," in Robert J. Gordon, ed., *The American Business Cycle: Continuity and Change*, Chicago: University of Chicago Press, 1986, pp. 395–458.

Friedman, Benjamin M. and Kuttner, Kenneth, "Money, Income, Prices, and Interest Rates," *American Economic Review*, June 1992, *82*, 472–92.

_____ **and** _____, "Why Does the Paper–Bill Spread Predict Real Economy Activity?" in James H. Stock and Mark W. Watson, eds., *New Research in Business Cycles: Indicators and Forecasting*, Chicago: University of Chicago Press, 1993 (forthcoming).

Gilson, Stuart, John, Kose and Lang, Larry, "Troubled Debt Restructurings: An Empirical Study of Private Reorganizations of Firms in Default," *Journal of Financial Economics*, October 1990, *27*, 315–53.

Gordon, Robert J. and Veitch, John M., "Fixed Investment in the American Business Cycle, 1919–83," in Robert J. Gordon, ed., *The American Business Cycle: Continuity and Change*, Chicago: University of Chicago Press, 1986, pp. 267–357.

Hoshi, Takeo, Kashyap, Anil, K and Scharfstein, David, "The Role of Banks in Reducing the Costs of Financial Distress in Japan," *Journal of Financial Economics*, September 1990, *27*, 67–88.

_____, _____ **and** _____, "Corporate Structure, Liquidity and Investment: Evidence from Japanese Industrial Groups," *Quarterly Journal of Economics*, February 1991, *106*, 33–60.

Kashyap, Anil K, Lamont, Owen and Stein, Jeremy, "Credit Conditions and the Cyclical Behavior of Inventories: A Case Study of the 1981–82 Recession," unpublished manuscript, Massachusetts Institute of Technology, 1992.

King, Stephen R., "Monetary Transmission: Through Bank Loans or Bank Liabilities?" *Journal of Money, Credit, and Banking*, August 1986, *18*, 290–303.

Maccini, Louis S. and Rossana, Robert, "Investment in Finished Goods Inventories: An Analysis of Adjustment Speeds," *American Economic Review*, March 1981, *71*, 218–36.

Meltzer, Allan, "Mercantile Credit, Monetary

Policy, and Size of Firms," *Review of Economics and Statistics*, November 1960, *42*, 429–37.

Oliner, Stephen D., Rudebusch, Glenn D., and Sichel, Daniel, "New and Old Models of Investment: A Comparison of Forecasting Performance," mimeo, Federal Reserve Board, Washington, DC, 1992.

Romer, Christina D. and Romer, David H., "New Evidence on the Monetary Transmission Mechanism" *Brookings Papers on Economic Activity*, 1990, (1), 149–213.

Rowe, Timothy, "Commercial Paper," in Timothy Q. Cook and Timothy D. Rowe, eds., *Instruments of the Money Market*, Richmond, VA: Federal Reserve Bank of Richmond, 1986, pp. 111–25.

Stiglitz, Joseph and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

Stock, James H. and Watson, Mark W., "New Indexes of Coincident and Leading Economic Indicators," *NBER Macroeconomics Annual*, 1989, *4*, 351–94.

Tobin, James and Brainard, William, "Financial Intermediaries and the Effectiveness of Monetary Control," *American Economic Review*, June 1963, *53*, 383–400.

Wooldridge, Jeffrey, "Notes on Regression with Integrated Processes," mimeo, Massachusetts Institute of Technology, 1990.

Board of Governors of the Federal Reserve System, *Introduction to Flow of Funds*, Washington, DC: Board of Governors of the Federal Reserve System, 1980.

Strategic Discipline in Monetary Policy with Private Information: Optimal Targeting Horizons

By MICHELLE R. GARFINKEL AND SEONGHWAN OH*

This paper analyzes a multiperiod monetary targeting procedure as a possible resolution to the credibility problem in policy when the monetary authority has some private information. By limiting the degree of flexibility permitted in policy, this procedure mitigates the credibility problem. As the length of the targeting horizon decreases, the severity of the credibility problem falls, but at the expense of weakening the monetary authority's ability to pursue its stabilization goals. Based on model simulations, the analysis studies the determinants of the optimal targeting horizon that balances the benefits of flexibility and discipline in policy. (*JEL* E50, F61)

How much flexibility or discretion should be given to the monetary authority? The answer to this time-honored question, in the monetarist spirit, is essentially none. Under the presumption that the monetary authority lacks an ability to forecast accurately or is able to forecast, at best, as well as economic agents, discretionary policy only creates an additional element of uncertainty that unnecessarily complicates economic agents' decision problems. Hence, the monetary authority should follow a strict rule, for example, a constant-money-growth-rate rule.[1] Against the monetarist prescription, proponents of an activist rule for the monetary authority argue that the inflexibility of a strict rule precludes an optimal response to unanticipated disturbances.

To this already complicated question, recent developments in the literature have added more complexity, highlighting the strategic elements of monetary policy. When the market-determined output or employment level is deemed suboptimal, a benevolent monetary authority might try to raise employment or output by surprising agents with high inflation if temporary nominal rigidities are present. Rational, forward-looking agents, recognizing this incentive, set high rates of wage inflation to discourage the monetary authority from trying to reduce their real wage below their target level. Hence, even if the monetary authority is perfectly benevolent in that it maximizes

*Garfinkel: Department of Economics, University of California-Irvine, Irvine, CA 92717; Oh: Department of Economics, University of California-Los Angeles, Los Angeles, CA 90024, and Department of Economics, Seoul National University, Seoul, 151-742, Korea. The first draft of this paper was completed while the first author was an economist and the second author was a visiting scholar at the Federal Reserve Bank of St. Louis. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System. We thank, without implicating, Gary Hansen, Gregory Hess, Axel Leijonhufvud, participants in the UCLA Money Workshop, and especially Guido Tabellini and two anonymous referees for their invaluable comments on earlier drafts of this paper. Scott Leitz provided useful technical assistance.

[1] See Allan H. Meltzer (1990), whose skepticism about the Fed's forecasting ability leads him to advocate a constant-money-growth-rate rule. Axel Leijonhufvud (1984) provides a detailed discussion about the uncertainty generated by discretionary policy under different monetary regimes.

social welfare, its policy might be inefficient; the equilibrium could be characterized by excessively high average inflation—an *inflationary bias*. As recognized since Finn Kydland and Edward Prescott's (1977) seminal paper, the efficient policy that avoids this bias might not be dynamically consistent or, more precisely, incentive-compatible.

Generally, when there is complete information, in the sense that individuals know the monetary authority's preferences and observe the realizations of the stochastic variables that constrain its choices, the reputational mechanism can serve to eliminate or, at least, to diminish the extent of the suboptimality arising from this dynamic consistency problem (see Robert J. Barro and David Gordon, 1983; Kenneth Rogoff, 1987).[2] If, however, the monetary authority has some private information, the efficacy of the reputational mechanism is called into question. Specifically, as shown by Matthew B. Canzoneri (1985), when the information structure is incomplete so that individuals cannot verify that the monetary authority has not intentionally invalidated their expectations, the efficacy of reputational considerations to resolve the credibility problem is weakened.[3] Moreover, private information precludes the effectiveness of any commitment technology to force the monetary authority to adhere to the optimal rule unless there is a separate mechanism to force the monetary authority to reveal its private information truthfully.

Given these difficulties, Canzoneri (1985) suggests taking a *legislative approach* in order to resolve the credibility problem when the policymaker has private information.[4] This approach assumes an incomplete commitment technology that cannot force the monetary authority to reveal its private information truthfully. Nevertheless, since a legislated procedure can be specified in terms of variables that are observed by all market participants (e.g., the growth rates of the monetary aggregates), this approach is potentially attractive. One example, studied by Canzoneri, is a two-period targeting procedure requiring that the average money growth rate per two periods equal the socially desired rate. More generally, provided that the procedure does not depend on the monetary authority's private information and the monetary authority can control perfectly the targeted variable, the procedure is operational.

This paper extends Canzoneri's (1985) analysis of a two-period targeting procedure in a multiperiod setting to investigate the scope of flexibility, in terms of the number of periods in the targeting horizon, that should be given to the monetary authority facing a trade-off between its output and inflation objectives. Legislation passed by Congress, for example, could mandate that the monetary authority announce a targeting horizon, over which time the growth of the money stock must average the rate corresponding to the (given) socially optimal inflation rate. Such an extension permits the analysis to identify the key factors that influence the optimal choice of the targeting horizon which balances credibility and flex-

[2]For a general discussion about dynamic inconsistency, see Brian Hillier and James Malcomson (1984). A necessary (but not sufficient) condition for dynamic inconsistency is that the government have fewer instruments than objectives.

[3]Following Edward J. Green and Robert H. Porter (1984), however, Canzoneri (1985) shows how reputational considerations can mitigate the credibility problem partially, even in the presence of private information. The solution rests on a trigger strategy that does not require explicit verification of the monetary authority's private information. However, this resolution is not complete. While the monetary authority adheres to the "reputational policy," excessively high rates of inflation would emerge on occasion. This solution, then, generates an expected rate of inflation in excess of the socially desired rate.

[4]For complementary analyses of monetary policy with private information, see Alex Cukierman and Meltzer (1986) and Jeremy C. Stein (1989). In contrast to these analyses, which mainly focus on maintained secrecy in monetary policy arising from private information, the present paper tries to specify an operational rule that improves upon the deadlock situation arising from the credibility problem.

ibility in monetary policy when there is private information.[5]

Under this procedure, the longer the targeting horizon, the greater is the degree of flexibility permitted in policy. However, as the targeting horizon increases, the greater is the monetary authority's willingness to act on its incentive to create surprise inflation in each period, and thus the greater is the (gross) inflationary bias in each period. In addition, the overall targeting constraint requires that these biases be reversed in subsequent periods. As the targeting horizon increases, then, the magnitude of those required reversals also increases. Thus, as shown below, the contribution of the remaining credibility problem to the variance in inflation around the socially optimal rate increases as the targeting horizon increases. The optimal length of the targeting horizon embodies the optimal trade-off between the cost of the remaining credibility problem and the cost of limiting flexibility to accommodate the predicted component of money demand shocks (i.e., the monetary authority's private information) relative to the policy that is fully discretionary.

As the monetary authority attaches greater importance to its objective for inflation relative to its objective for output, the optimal targeting horizon becomes longer. The basic intuition for this result is that the monetary authority's incentive to create surprises decreases, thereby decreasing the severity of the credibility problem and, hence, increasing the possible benefits of flexibility afforded by the N-period targeting procedure to permit the monetary authority to react to its private information. For the extreme case in which the inflationary bias that emerges in the full-discretionary regime is so small that the constraint imposed by the targeting procedure becomes unnecessary, the optimal number of periods in the targeting horizon is infinite. For the other extreme case, in which the credibility problem is sufficiently severe to make the optimal targeting horizon equal to one period, the procedure simplifies to a constant-money-growth rule. Such a rule eliminates the equilibrium inflationary bias associated with the credibility problem, but at the cost of removing all flexibility in policy. Nevertheless, as the weight that the monetary authority attaches to its inflation objective relative to its output objective approaches zero, the constant-money-growth rule is more likely to emerge as the optimal targeting policy.

In what follows, the next section briefly presents a model of monetary policy with private information in a multiperiod setting and analyzes the monetary authority's optimal, time-consistent monetary policy subject to the N-period average targeting constraint. Section II presents simulations of the model to illustrate how the parameters of the model influence the optimal choice of

[5]Other existing analyses of institutional arrangements that can possibly mitigate the inefficiencies associated with the credibility problem (e.g., Rogoff, 1985; Susanne Lohmann, 1992) similarly emphasize the trade-off between credibility and flexibility. However, they abstract from the complications that can arise when the monetary authority has (nonverifiable) private information. Indeed, as argued by Canzoneri (1985), Rogoff's (1985) *perverse* policymaker solution to the credibility problem would not be effective when the monetary authority has private information concerning its own preferences as well as the realizations of the stochastic variables that constrain its choices. This criticism would apply as well to Lohmann (1992), who combines Rogoff's (1985) perverse policymaker solution with Robert Flood and Peter Isard's (1989) mixed-strategy solution. In addition, as will be discussed, these analyses focus on flexibility within a period, rather than over a number of periods, resulting in a slightly different (but complementary) notion of flexibility from that developed here. In any case, because the institutional solutions (including the one analyzed in this paper) involve limits on policy, they are not complete resolutions to the credibility problem. Generally, whether the institutional approach or the reputational approach provides a more effective resolution would depend, in practice, on the monetary authority's ability to influence expectations relative to its ability to design and establish institutions. Answering this question is difficult at best. Nevertheless, it should be feasible to derive the conditions under which any one of the institutional solutions produces a smaller deadweight loss than that associated with Canzoneri's (1985) reputational solution. Although the present analysis does not analytically derive those conditions, a numerical example in which the reputational solution is clearly dominated by the targeting procedure studied here is presented (see footnote 16).

the targeting horizon. These simulations suggest that, for many parameter values, the cost of limiting flexibility to reduce the severity of the credibility problem will be too large to render a finite-period targeting procedure desirable; but, for many other parameter values, some discipline in policy will be warranted, even if the monetary authority's forecasting ability is fairly accurate. Indeed, for a large part of the parameter space, the constant-money-growth rule emerges as the optimal policy. The present analysis suggests, then, that the credibility problem, as well as the monetary authority's forecasting ability, plays an important role in the optimal design of feasible monetary rules. Finally, Section III offers some concluding remarks, including possible extensions of the analysis.

I. A Model of Monetary Policy with Multiperiod Targeting

This section presents a simple economic model to investigate the efficacy of multiperiod average targeting procedures to support a better outcome than that achieved with full discretion when the monetary authority has private information. Following conventional practice, the analysis builds on a standard rational-expectations supply function,

$$(1) \qquad y_t = y^n + \theta(\pi_t - \pi_t^e) \qquad \theta > 0$$

where y_t denotes the logarithm of output in period t, y^n denotes the logarithm of the natural level of output, π_t is the inflation rate in period t, and π_t^e is the wage setters' expection of inflation conditional on information available at the end of period $t-1$.[6]

The following money demand equation determines the equilibrium price level:

$$(2) \qquad m_t - p_t = y^n - v_t$$

where m_t and p_t equal, respectively, the logarithms of the money stock and the price level in period t and v_t is a money demand disturbance realized at the end of period t. The disturbance is assumed to follow a random walk. The equilibrium inflation rate is obtained by taking the first difference of (2),

$$(3) \qquad g_t - \pi_t = \delta_t$$

where $\delta_t = v_{t-1} - v_t$ and g_t is the growth rate of money, the monetary authority's instrument;[7] δ_t is an independently and identically distributed random variable with a zero mean and a finite variance, σ_δ^2. When wages are set, δ_t is not known (i.e., $\delta_t^e = 0$). The expression in (3), then, implies that wage setters' expectation of inflation depends solely on the monetary authority's policy, g_t. This expectation, g_t^e, ultimately depends on the monetary authority's strategy.

[6]The emergence of the inflationary bias, discussed later in the paper, rests on the assumption embedded in this equation that workers supply labor along the demand schedule. However, as Cukierman (1992 Ch. 3) argues, it is highly unlikely that employment is determined along the labor demand curve when labor demand is in excess of supply. Cukierman demonstrates that, alternatively, when employment is viewed as being determined by the "short side" of the market, the monetary authority's incentive to create surprise inflation disappears and so does the inflationary bias. Nevertheless, some degree of unionization that keeps real wages *above* the real market-clearing wage can give the monetary authority an incentive to create surprise inflation to reduce (unexpectedly) the real wage, because the union's best strategy is to encourage its members to supply labor *ex post* along their individual labor supply curves. See Cukierman (1992) for the conditions under which the presence of unions gives rise to an inflationary bias.

[7]As indicated earlier, for the targeting procedure to work, the analysis must assume that the monetary authority can control the growth rate of money perfectly if that is its target. That the Federal Reserve System chooses to target the more broadly defined monetary aggregates (i.e., M2 and M3) rather than those over which it has more control (i.e., high-powered money and M1) could be related to the secrecy problem. See, for example, Cukierman and Meltzer (1986) who show that the monetary authority might choose a less precise procedure for policy implementation so as to maintain some degree of ambiguity, whereby the secrecy of its information could be maintained to some extent. The present analysis, however, does not permit meaningful distinctions among the various monetary aggregates.

To study alternative strategies for the monetary authority, the analysis to follow assumes that the monetary authority chooses its policy to maximize its expected N-period average utility:

$$(4)\quad U' = E_0\left\{\frac{1}{N}\sum_{t=1}^{N} u'_t\right\} \qquad 1 \le N < \infty$$

where one-period utility in period t is given by

$$u'_t = -(y_t - ky^n)^2 - s(\pi_t - \pi^*)^2 \qquad k > 1.$$

$E_\tau\{\cdot\}$ is an expectations operator conditional on period $t = \tau$ information, and N denotes the number of periods in the targeting horizon.[8] For $N = \infty$, the expected average utility is given by $E_0\{u'_t\}$. The logarithm of the socially desirable output level is represented as ky^n, and π^* is the socially determined optimal rate of inflation; s is the weight the monetary authority attaches to its goal of inflation stability relative to its goal of output stability.[9]

Letting $U \equiv U'/\theta^2$ and $u_t \equiv u'_t/\theta^2$ and using (1) and (3) with the assumption that $\delta_t^e = 0$, the monetary authority's expected average N-period utility can be rewritten as

$$(5)\quad U = E_0\left\{\frac{1}{N}\sum_{t=1}^{N} u_t\right\}$$

where

$$u_t = -(g_t - g_t^e - \delta_t - y^*)^2 - f(g_t - \delta_t - \pi^*)^2$$

and $f \equiv (s/\theta^2)$ and $y^* \equiv (k-1)y^n/\theta$.

After period-t wages are set, the monetary authority chooses its policy, g_t. In contrast to wage-setters, the monetary authority receives some information about the disturbance to money demand (δ_t) after wages are set but before policy actions are taken. Specifically, it has a *private* forecast of this disturbance, $d_t = E_t\{\delta_t\}$, that satisfies

$$(6)\quad \delta_t = d_t + \varepsilon_t$$

where ε_t is an independently and identically distributed disturbance realized after policy is implemented. This forecast error has a zero mean, a finite variance, σ_ε^2, and no correlation with d_t. Similarly, d_t is independently and identically distributed with a zero mean and a finite variance, σ_d^2. Thus, $\sigma_\delta^2 = \sigma_d^2 + \sigma_\varepsilon^2$. Section II discusses the case when d_t is serially correlated. Although wage-setters observe δ_t and π_t after g_t is set, they cannot distinguish the monetary authority's forecast, d_t, from the forecast error, ε_t.[10]

[8] N can be thought of as the time horizon of an individual policymaker. However, given the stationary nature of the model, the solution would not change if, instead, the policymaker were infinitely-lived since the infinite-horizon problem can be decomposed into independent N-period-horizon problems. The implicit assumption here, that the monetary authority does not discount the future, is not strong. A discount factor less than 1 only complicates the analysis without providing much additional insight. (A copy of the solution in the case of discounting is available from the authors upon request, and the effects of discounting on the optimal targeting horizon are discussed briefly at the end of Section II.)

[9] See Canzoneri (1985 pp. 1058–9) for a detailed discussion of the possible interpretations of the assumption that $k > 1$, an assumption that is important for explaining the emergence of a positive inflationary bias. These interpretations build on distortions (presumed to exist in the economy) that depress average output, or the natural level of output, below the level considered to be optimal from a social-welfare perspective (but see footnote 6 and Cukierman [1992 Ch. 3], who provides a useful critique of this social-welfare approach). As an alternative, Cukierman (1992) adopts a political approach to study central-bank behavior. This approach interprets the monetary authority's output objective with $k > 1$ as representing those interest groups advocating economic stimulation. Under this interpretation, the parameter s reflects "a compromise reached through the central bank between the advocates of economic stimulation and the advocates of price stability" (Cukierman, 1992 p. 43). This approach would seem inappropriate for the present representative-agent model, however.

[10] Given the sequence of actions by wage-setters and the monetary authority relative to the timing of the realization of the monetary authority's private informa-

Before the optimal, time-consistent monetary policy with multiperiod targeting is derived, two benchmark solutions are presented. The first solution, the efficient solution, is that which would obtain if there were a full commitment technology permitting the monetary authority to adhere to a contingent rule, while truthfully revealing its private information. The second solution, the full-discretionary solution, assumes that the monetary authority takes wage-setters' expectations as given. It is identical to the solution that would emerge under an infinite targeting procedure, since $N=\infty$ is equivalent to no constraints on policy.

A. *The Efficient Solution*

Assuming a full commitment technology to force the monetary authority to reveal its private information truthfully and to adhere to a contingent rule so that it can influence wage-setters' expectations effectively in a way that is consistent with its policy, the optimal monetary policy, $\hat{g}$, is given by

$$\hat{g}_t = \pi^* + d_t \tag{7}$$

for $t=1,2,\ldots$, which yields the following expected average utility for the monetary authority:

$$\hat{U} = -(1+f)\sigma_\varepsilon^2 - (y^*)^2. \tag{8}$$

Note that the money growth rule in (7) completely accommodates money demand shocks to stabilize inflation around the socially desired rate.

B. *The Full-Discretionary Solution*[11]

Under an alternative assumption that the monetary authority behaves *as if* it can have no impact on wage-setters' expectations, it maximizes the expectation of (5) conditional on its forecast of the disturbance to money demand, subject to (6) and taking g_t^e as given. The wage-setters' expectation of the associated first-order condition implies $g_t^e = \pi^* + y^*/f$. By substituting g_t^e back into the first-order condition, one can verify that the money growth rate under this regime, $\bar{g}$, is given by

$$\bar{g}_t = \pi^* + d_t + \frac{y^*}{f}. \tag{9}$$

Note that the monetary authority's private information is fully revealed in the outcome. Although the monetary authority can do no better than to set its policy according to (9) given wage-setters' expectations, the monetary authority's incentive to create surprise inflation generates an inflationary bias, y^*/f. Even if the monetary authority had no private information, this bias would emerge, provided that it took expectations as given.

The inefficiency of the full-discretionary solution is revealed by comparing the expected average utility of the monetary authority in this regime, given by

$$\bar{U} = -(1+f)\sigma_\varepsilon^2 - \left(1+\frac{1}{f}\right)(y^*)^2 \tag{10}$$

to that obtained in the efficient regime, given by (8). The difference between (8) and

tion and the transitory nature of that private information, Vincent Crawford and Joel Sobel's (1982) notion of *cheap talk*, recently applied to the monetary-policy problem by Stein (1989), is not operational in the context of this model. (See Oh and Garfinkel [1990] for a brief discussion of why the two-dimensionality of the credibility problem in monetary policy with private information obscures the relevance of the cheap-talk mechanism. To make cheap talk effective in this framework with a slight variation in the sequence of events, it is necessary to limit flexibility in monetary policy.) Section III discusses one extension of the model to incorporate the possibility of cheap talk in the multiperiod targeting procedure. It should be noted here that the assumption that $\delta_t^e = 0$ is not crucial. Even if wage-setters had reasonable forecasts of δ_t, the analysis to follow would be relevant, provided that the monetary authority's forecast is private.

[11] In the static framework, this solution is equivalent to the (one-shot) Nash solution. As already noted and as discussed in part C of this section, however, this solution more generally is interpreted as an infinite-period, average targeting procedure.

(10), $(y^*)^2/f$, captures the disutility of the inflationary bias.

C. *The N-Period Targeting Solution*

As is widely recognized, while the full-discretionary policy is not first-best, the efficient solution is not necessarily a feasible outcome in the absence of a full commitment technology. That is, the monetary policy in (7) is not incentive-compatible. Once wages are set, the monetary authority's optimal policy no longer corresponds to (7). Rather, the monetary authority would like to set $g_t = \pi^* + d_t + y^*/(1+f)$. In addition, the monetary authority's private information obscures the role of reputational considerations to diminish the inflationary bias. Specifically, without a complete information structure, wage-setters cannot verify that the monetary authority has followed the efficient policy.[12] Moreover, the monetary authority's private information precludes the effectiveness of legislation requiring the monetary authority to implement policy according to (7). Even if the legislation were binding, the monetary authority would have an incentive to lie, claiming that its forecast of δ_t equaled $d_t + y^*/(1+f)$, whereby it could disguise the optimal cheating policy (given $g_t^e = \pi^*$) as the efficient one. Without any additional constraints on policy, the presence of nonverifiable private information means that the monetary authority cannot influence wage-setters' expectations effectively, and the full-discretionary policy emerges in the equilibrium outcome.

As a possible resolution to the inefficiency of the inflationary bias when the monetary authority cannot credibly reveal its private information, Canzoneri (1985) studies a two-period average targeting procedure, requiring $g_1 + g_2 = 2\pi^*$. This procedure is attractive not only in its simplicity, but in its independence from the monetary authority's private forecast. Such independence makes it operational, for enforcement of the (precommitted) constraint does not require verification of d_t.[13] As a generalization of Canzoneri's analysis, consider an extended average targeting procedure that requires

$$(11) \qquad \sum_{t=1}^{N} g_t = \pi^* N$$

where $1 \le N < \infty$. For $N = \infty$, there is no relevant constraint on the conduct of policy.

To derive the optimal, dynamically consistent monetary policy under this (precommitted) targeting regime, a backward solution concept is appropriate. Specifically, in period $t = N-1$, the monetary authority maximizes the sum of the expected value of (5) in $t = N$ and $t = N-1$ with respect to g_{N-1} subject to the forecast d_{N-1} and (11). In solving this optimization problem, the monetary authority takes g_N^e and g_{N-1}^e as given while recognizing the impact that g_{N-1} will have on future expectations, g_N^e, through the constraint in (11). Repeating this exercise for $t = N-2, N-3, \ldots$, one can find that

$$(12) \quad \tilde{g}_t = \tilde{g}_t^e + \frac{(N-t)(1+f)}{N-t+(N-t+1)f} d_t$$

for $t = 1, 2, \ldots, N$, where $\tilde{g}$ denotes the

[12] But see footnote 3.

[13] To be sure, the analysis to follow implicitly assumes that a precommitment to adhere to the targeting constraint in (11) is credible. Even if the monetary authority's expected average utility under this procedure (for a given N) exceeds that which would be obtained under the full-discretionary solution ($N = \infty$), the constraint need not be incentive-compatible in any period during the targeting horizon. In other words, in any period $t \le N < \infty$, given expectations, the monetary authority has an incentive to violate the constraint. This incentive undermines the credibility of the precommitment. The present analysis implicitly assumes, then, that there exists an external enforcement mechanism (e.g., legislation effectively requiring that the monetary authority be severely punished if it fails to hit that average period target over the prespecified horizon, N) to support the procedure provided that it yields a better outcome than that obtained under full discretion, a condition that is examined in Section II. Note that, when this condition is satisfied, reputational considerations might be able to substitute for the assumed commitment technology to support the targeting rule in equilibrium.

money growth policy under the N-period targeting procedure and

$$\tilde{g}_t^e = \frac{N\pi^*}{N-t+1} + \frac{(N-t)y^*}{(N-t+1)f} - \sum_{\tau=1}^{t-1} \frac{\tilde{g}_\tau}{N-t+1}$$

for $t = 1,2,\ldots,N$. Note that, in the expression above, the summation is taken only for $t > 1$. (In what follows, Σ_s^t is taken to be equal to zero whenever $s > t$.) The Appendix illustrates the solution method used here more explicitly.

To eliminate past $\tilde{g}$ so as to express the optimal policy only in terms of current and past d as well as the parameters of the model, (12) is used to find $\tilde{g}_1$ and sequentially used to find $\tilde{g}_t$, $t = 2,3,\ldots,N$, by substituting in past $\tilde{g}_t$. That is,

$$\tilde{g}_t = \Gamma(t) - \sum_{\tau=1}^{t-1} \frac{\Gamma(\tau)}{N-\tau} \tag{13}$$

where

$$\Gamma(t) = \frac{N\pi^*}{N-t+1} + \frac{(N-t)y^*}{(N-t+1)f} + \frac{(N-t)(1+f)}{N-t+(N-t+1)f} d_t$$

for $t = 1,\ldots,N$. By rearranging (13) and simplifying, one can verify that the optimal monetary policy subject to the N-period targeting constraint is given by the following:

$$\tilde{g}_t = \pi^* + \frac{(N-t)y^*}{(N-t+1)f} + \frac{(N-t)(1+f)}{N-t+(N-t+1)f} d_t - \sum_{\tau=1}^{t-1} \left[\frac{y^*}{(N-\tau+1)f} + \frac{1+f}{N-\tau+(N-\tau+1)f} d_\tau \right] \tag{14}$$

for $t = 1,2,\ldots,N$.

Expected average utility under this regime is given by

$$\tilde{U} = -(y^*)^2 - (1+f)\sigma_\varepsilon^2 - \frac{(y^*)^2}{Nf} \sum_{t=1}^{N} \left[\frac{N-t}{N-t+1} - \sum_{\tau=1}^{t-1} \frac{1}{N-\tau+1} \right]^2 - \frac{f^2(1+f)}{N} \sigma_d^2 \sum_{t=1}^{N} \left[\frac{1}{N-t+(N-t+1)f} \right]^2 - \frac{f(1+f)^2}{N} \sigma_d^2 \sum_{t=2}^{N} \sum_{\tau=1}^{t-1} \left[\frac{1}{N-\tau+(N-\tau+1)f} \right]^2. \tag{15}$$

Note that the sum of the first two terms in the above expression equals the expected average utility that would be realized if the efficient outcome could be supported in equilibrium, (8). Hence, the last three terms reflect the inefficiency of the targeting procedure.[14]

[14] For $N = 1$, equation (15) reduces to the average expected utility under a constant-money-growth rule, given by $-(y^*)^2 - (1+f)\sigma_\delta^2$. In this case, the last three terms reduce to $(1+f)\sigma_d^2$, the cost arising from removing all flexibility in policy. As N approaches infinity, (15) reduces to (10), the average expected utility under the full-discretionary regime. To verify this, note that the third term approaches $(y^*)^2/f$ as N approaches infinity, while the last two terms converge to zero. The convergence of the last term to zero can be verified easily by observing that the double summation of that term can be written as

$$\frac{1}{N} \sum_{t=1}^{N} \frac{t-1}{(t-1+tf)^2}.$$

Since the limit of $(t-1)/(t-1+tf)^2$ converges to zero as t approaches infinity, the arithmetic mean of such a sequence will also be zero (see Walter Rudin, 1976 p. 80). Thus, when $N = \infty$, the last three terms in (15) boil down to $(y^*)^2/f$, the disutility of the inflation that would emerge in the full-discretionary solution.

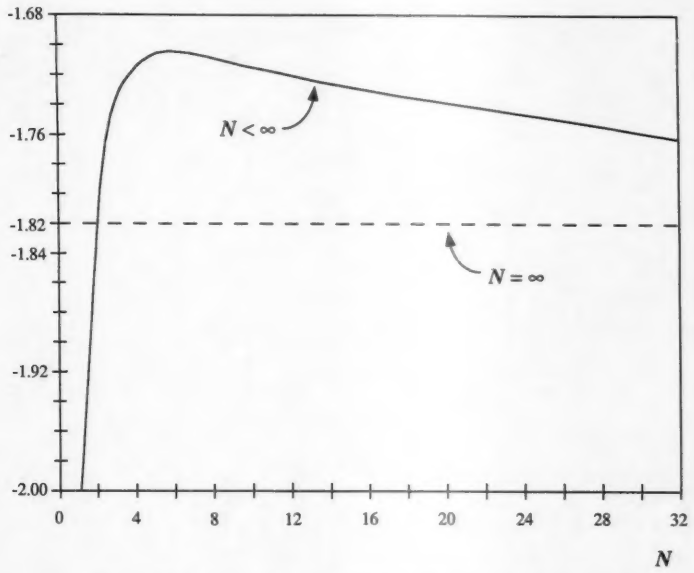


FIGURE 1. NORMALIZED INEFFICIENCY OF MULTIPERIOD TARGETING PROCEDURES

The optimal targeting horizon, N^*, maximizes (15). Assuming that N^* is finite,[15] the average inefficiency is increasing in N for $N > N^*$, as shown with an example in Figure 1. The horizontal line in the figure measures the normalized inefficiency of the full-discretionary solution when $f = 1.0$ and $y^*/\sigma_d = 1.35$ [i.e., $(y^*)^2/f\sigma_d^2 = 1.82$]. The inverted V-shaped curve measures the normalized, average inefficiency of a multiperiod targeting over various N, showing that $N^* = 6$ for the chosen parameter values.[16]

Given the N-period targeting constraint, the monetary policy in (14) is dynamically consistent, like the full-discretionary solution without any constraints. Note, however, that for finite N the *gross* inflationary bias [i.e., the second term in (14)] is lower than that in the full-discretionary solution. From (13) or (14) in $t = 1$, for example, one can verify that the bias is $(N-1)y^*/Nf$, which is reversed in each of the subsequent $N-1$ periods, in increments of y^*/Nf. The increments are of the same magnitude since the monetary authority does not discount the future. Similarly, the bias in $t = 2$ is $(N-2)y^*/(N-1)f$, which is reversed in the subsequent $N-2$ periods in increments of $y^*/(N-1)f$. As t approaches N, the monetary authority's willingness to act on its incentive to create surprise inflation falls, and so does the gross inflationary bias. At the same time, there is an increasing accumulation of past biases to be reversed. Thus, as t approaches N, the period-t inflationary bias net of the sum of current reversals of previous periods' biases (referred to here as the *remaining* or *net* inflationary bias) falls and eventually becomes negative to make

[15]Although writing down a closed-form solution for N^* is extremely difficult, it is possible to verify that N^* can approach infinity for a certain set of parameter specifications (see Section II).

[16]It should be noted that, for $\sigma_d < 1.5$, Canzoneri's (1985) reputational solution would not be feasible since it requires that $y^*/f > 2$ (p. 1065). For this numerical example, then, the reputational solution is clearly dominated by the institutional solution studied in this paper, when $\sigma_d < 1.5$.

TABLE 1—THE NET INFLATIONARY BIASES OVER N

	t										
N	1	2	3	4	5	6	7	8	9	10	b
1	0.00										0.00
2	0.39	−0.39									0.15
3	0.52	0.13	−0.65								0.24
4	0.59	0.33	−0.07	−0.85							0.29
5	0.62	0.43	0.17	−0.22	−1.00						0.33
6	0.65	0.49	0.30	0.04	−0.35	−1.13					0.36
7	0.67	0.54	0.38	0.19	−0.07	−0.46	−1.24				0.38
8	0.68	0.57	0.44	0.29	0.09	−0.17	−0.56	−1.34			0.40
9	0.69	0.60	0.48	0.35	0.20	0.00	−0.26	−0.65	−1.43		0.42
10	0.70	0.62	0.52	0.41	0.28	0.12	−0.08	−0.34	−0.73	−1.51	0.43

Notes: In this example, $f = 1$, $y^* = 0.78$, and $\pi^* = 0.02$. The inflationary bias when $N = \infty$ equals 0.78. For this example $N^* = 6$. For the last column, b denotes the disutility from the remaining inflationary bias [i.e., the absolute value of the third term in equation (15)]. Note that the rows might not add to zero due to rounding.

the average remaining inflationary bias over the targeting horizon zero.

The movement in the remaining inflationary bias (i.e., the unconditional expectation of money growth in excess of the socially determined optimal inflation rate) over t is given in Table 1 for $1 \leq N \leq 10$. The table is constructed on the basis of the previous example with the additional assumptions that $\pi^* = 0.02$ and $y^* = 0.78$ (implying that $\sigma_d^2 = 1/3$). For this example, the inflationary bias that would emerge in the full-discretionary regime is 0.78 for all t. As revealed by the table, although the remaining inflationary bias averages zero over the horizon, this remaining bias in *absolute* terms can exceed the bias that would emerge in the full-discretionary solution. In addition, the table indicates that increases in N imply a greater average of the absolute value of the remaining inflationary biases that emerge over the targeting horizon due to the constraint (11). The basic idea here is similar to that behind the downward movement of the remaining inflationary bias over the horizon. As N increases, the targeting constraint (11) appears less binding to the monetary authority at any given period t. Accordingly, the monetary authority is more inclined to act on its incentive to create surprise inflation, which is recognized by wage-setters. Of course, the future reversals must be greater, in magnitude, if the targeting constraint (11) is to be satisfied. Thus, as N increases, the variance of the remaining bias around zero increases. However, because the remaining biases are perfectly anticipated by wage-setters, the increased variance of the remaining bias has no implications for the variation of output.

The disutility of the remaining inflationary biases is captured by the third term in (15). For $N = 1$, this term equals zero. Because any period's (gross) inflationary bias (which must eventually be reversed over the remainder of the targeting horizon) increases as N increases, the loss in utility due to the effects of the remaining inflationary bias on the variability of inflation around π^* increases as N becomes large. As N goes to infinity, this loss converges to $(y^*)^2/f$; in the example above, this loss, reported in the last column of Table 1, converges to 0.608.

In contrast to the full-discretionary and the efficient solutions, the monetary policy in (14) involves only a partial accommodation of the current shock. The targeting procedure limits the monetary authority's flexibility to react to its forecast. However, as N approaches infinity, the accommodation is full. The effect of increasing the monetary authority's flexibility to react to its private information is partly captured by the fourth term in (15), which equals $(1 + f)\sigma_d^2$ for $N = 1$ and is decreasing in N, approaching zero as N approaches infinity. That is, by relaxing the monetary authority's con-

straint, increasing the length of the targeting horizon provides more leeway for the monetary authority to achieve its output and inflation-stabilization goals.

The constraint in (11), however, requires that the accommodation in every period be reversed in subsequent periods. The effect of this requirement can be observed easily in (13) or (14). In particular, as the constraint becomes more binding (i.e., t approaches N), the monetary authority's reaction to the current shock approaches zero. The reaction to the shock in time t, d_t, is reversed in increments of

$$\frac{1+f}{N-t+(N-t+1)f}d_t$$

in each of the remaining $N-t$ periods. Again, that the increments of the reversal of a given reaction to a time-t shock are of equal magnitude is driven by the assumption that the monetary authority does not discount the future. Note that these reversals have no implications on the variability of output. In this model, the optimal policy, given by (12), implies that past money disturbances, d_t, are fully revealed to wage-setters upon their observing g_t. Thus, wage-setters fully incorporate these reversals into expectations. However, the reversals do influence the variability of inflation around π^*.

The fifth term in (15) reflects the cost of the increased variability of inflation due to the requirement of having to reverse earlier reactions to shocks. As N increases from 1, that loss in expected utility initially increases from zero. That loss, however, decreases as N becomes sufficiently large. At this point, increasing flexibility reduces the expected cost of engineering those reversals, so that the loss captured by the fifth term falls and finally converges to zero. Combining the fourth and fifth terms of (15) yields the net effect of limiting the scope of flexibility as permitted by the targeting procedure.

That the sum of the last terms in (15) is increasing in magnitude in N for $N > N^*$ implies that the targeting procedure becomes more inefficient as N increases above the optimal targeting horizon. Nevertheless, note that a sufficient condition for N^* to be finite is that $(1+f)\sigma_d^2 < (y^*)^2/f$ (i.e., that a strict one-period targeting rule dominates the full-discretionary solution) as can be verified by comparing the expected average utility under the targeting procedure, given by (15) with $N=1$, to the expected average utility in the full-discretionary regime, given by (10) or by (15) with $N=\infty$. This condition is not necessary, however. More generally, as revealed by a comparison of (10) and (15), the targeting procedure will dominate the full-discretionary solution ($N=\infty$) provided that the magnitude of the sum of those terms, for any finite N, is less than $(y^*)^2/f$, the disutility of the inflationary bias that would emerge in the full-discretionary equilibrium.

The basic idea here is as follows: as N goes from 1 to infinity, the variance of output falls, approaching that which would obtain in the full-discretionary regime. At the same time, the variance of inflation around π^* due to the remaining credibility problem increases, approaching that which would obtain in the full-discretionary regime. Although increasing N also implies more flexibility to insulate inflation and output from money demand shocks, the required reversals of reactions to past d_t adds to the variability of inflation around π^*, at least initially as N increases from 1. The variability due to such reversals over the targeting horizon falls sometime thereafter, however, approaching zero as N goes to infinity. Thus, even if the strict one-period targeting procedure does not dominate the full-discretionary policy, some limits on flexibility might be desirable. For N^* to be finite under conditions for which the full-discretionary policy dominates the constant-money-growth rule, the disutility of the increased variability of inflation around π^* due to the reversals of earlier reactions to d_t must fall soon after N increases from 1 and approach zero at a rapid rate as N continues to increase. Otherwise, the reduction in the variability of inflation around the socially determined rate below that which obtains in the full-discretionary regime will

TABLE 2—OPTIMAL TARGETING HORIZONS, N^*

	f									
y^*/σ_d	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.1	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
0.2	55	∞	∞	∞	∞	∞	∞	∞	∞	∞
0.3	5	∞	∞	∞	∞	∞	∞	∞	∞	∞
0.4	2	9	∞	∞	∞	∞	∞	∞	∞	∞
0.5	1	3	13	∞	∞	∞	∞	∞	∞	∞
0.6	1	2	4	16	∞	∞	∞	∞	∞	∞
0.7	1	1	2	5	19	∞	∞	∞	∞	∞
0.8	1	1	1	3	6	22	∞	∞	∞	∞
0.9	1	1	1	2	3	7	25	∞	∞	∞
1.0	1	1	1	1	2	3	7	28	∞	∞
1.1	1	1	1	1	1	2	3	8	30	∞
1.2	1	1	1	1	1	1	2	4	8	32
1.3	1	1	1	1	1	1	1	2	4	9
1.4	1	1	1	1	1	1	1	2	2	4
1.5	1	1	1	1	1	1	1	1	2	2
1.6	1	1	1	1	1	1	1	1	1	2
1.7	1	1	1	1	1	1	1	1	1	1

Note: Strictly speaking, $N^* = \infty$ indicates that the full-discretionary solution dominates only those for finite multiperiod targeting procedures with $N \leq 200$.

not be sufficient to compensate for the increased variability in output induced by the targeting constraint; in this case, no constraints on policy through the targeting procedure would be desirable.

II. Optimal Targeting Horizons: Simulations of the Model

The optimal degree of flexibility depends on the accuracy of the monetary authority's forecasts of money demand disturbances, as measured by σ_d^2 for a given σ_δ^2 or a given σ_ε^2, and the severity of the credibility problem, as measured by y^*/f. As in much of this literature, however, analytically deriving the optimal targeting horizon, N^*, that maximizes expected average utility, as given by (15), is not possible despite the simplicity of the model. This section summarizes the results of simulations that specify the two key parameters of the model: y^*/σ_d (the weighted difference between the player's output goals relative to the magnitude of the predictable part of the shock) and f (the ratio of the weight attached to inflation stability relative to output stability in the monetary authority's preferences, s, to the squared elasticity of output with respect to unanticipated inflation, θ). For various parameter specifications, N^* is reported in Table 2. Discrete jumps in N^*, as the parameter values vary, are due to the fact that changes in the parameter values are not sufficiently small.

A close inspection of the table reveals two important implications of the simulation exercise. First, for a given y^*/σ_d, as f increases, the optimal targeting horizon gets longer. That is, as the weight of importance that the monetary authority attaches to its inflation objective relative to that for its output objective increases, the optimal length of the targeting horizon increases.[17]

[17]Note that an analogous comparative-static exercise involving a hypothetical decrease in the elasticity of output with respect to unanticipated inflation, θ, would not necessarily produce the same result. A smaller θ implies not only a larger f, but a larger y^*. As we will discuss, a larger y^*, for a given f, implies a smaller N^*. Thus, hypothetical changes in f should be interpreted as changes in s only. Similarly, hypothetical changes in y^* should be interpreted as changes in either k or y^n only.

The basic idea behind this implication is rather simple. As f increases, the monetary authority's incentive to push output beyond its natural level falls so that the inflationary bias that would emerge in the full-discretionary equilibrium falls. Accordingly, the optimal scope of flexibility under the targeting procedure increases. For large values of f, the optimal targeting solution boils down to the full-discretionary solution, in which there is complete flexibility. In this case, the inflationary bias that emerges under full discretion becomes so trivial that the limits on flexibility imposed by the targeting procedure are unnecessary.

This implication is similar to that of Rogoff's (1985) analysis. In contrast to the present analysis, however, Rogoff distinguishes the monetary authority's preferences from that of society and thereby gives an institutional interpretation to the monetary authority's preferences.[18] The difference emerges in the weight of importance attached to the inflation goal relative to the output goal; for convenience, the difference is denoted here by $\mu \equiv f_m - f > 0$, where f reflects society's preferences and f_m reflects the (perverse) monetary authority's preferences. Nevertheless, as in Rogoff's analysis, the inflationary bias that emerges in the full-discretionary outcome becomes less severe, as the monetary authority's dislike for deviations of actual inflation away from the socially determined optimal rate increases. The important distinction between the present analysis and that of Rogoff lies in his interpretation of flexibility. Specifically, in a one-shot-game setting, he roughly interprets increasing the monetary authority's relative concern for inflation stability around π^* (i.e., increasing f_m for a given f) as limiting flexibility, for it implies less complete and, for a given f, less efficient accommodations of *supply* shocks by the monetary authority.[19] In contrast to Rogoff (1985), the present multiperiod analysis interprets increases in flexibility in terms of increasing N, finding that the more *perverse* is the monetary authority (i.e., the larger is f) the greater are the advantages of full flexibility in monetary policy.

The second implication of the simulation exercise is that, for a given f, the optimal targeting horizon becomes shorter as y^*/σ_d increases. There are two potential forces at work here. As y^* increases for a given σ_d, the difference between the socially optimal output target and the natural level of output becomes larger; as a consequence, the inflationary bias that would emerge in the full-discretionary solution becomes larger, thereby detracting from the benefits of flexibility afforded by the targeting procedure. Furthermore, as σ_d falls for a given y^*, the expected value of reacting to current shocks (the monetary authority's private information) to stabilize inflation and output falls. Since $\sigma_d^2 = \sigma_\delta^2 - \sigma_\epsilon^2$, a decrease in σ_d^2 can be thought of as a decrease in σ_δ^2 or as an increase in σ_ϵ^2. In either case, the reduction in σ_d^2 implies a reduction in the degree of accuracy of the monetary authority's forecast, $\sigma_d^2/\sigma_\delta^2$. Not surprisingly, then, the simulation exercise suggests that, with a

[18]Also see Lohmann (1992). For a general survey on this and related issues, see Torsten Persson and Guido Tabellini (1990).

[19]Under this interpretation, then, the fully flexible policy is that which would otherwise be implemented by the policymaker with preferences corresponding exactly to those of society. In her interesting analysis of what incentive structure should be imposed on the monetary authority, Lohmann (1992) adds another dimension to this interpretation of flexibility with a notion of central-bank independence—in particular, independence of the monetary authority from the policymaker having preferences identical to those of society. Building on Flood and Isard (1989), her notion of independence revolves around the possibility that the policymaker can, at a cost, dismiss the monetary authority and implement the full-discretionary policy. In equilibrium, the monetary authority follows a policy involving a *flexible escape clause*, whereby it can avoid dismissal. The smaller the cost of dismissal, the smaller is the degree of independence and, for a given f_m, the closer is the monetary authority's reaction to the shock to what the policymaker's reaction would be if it were to formulate and otherwise implement policy directly. Thus, for a given f_m, a decrease in independence could be interpreted as an increase in flexibility. (Note that Rogoff's [1985] analysis implicitly assumes that the cost of dismissal is infinite, implying no independence.)

reduction in the monetary authority's forecasting ability, the optimal limits on flexibility are greater (i.e., N^* is smaller). The strict one-period targeting rule is more likely to be optimal as y^*/σ_d rises.

This implication further distinguishes the results of the present analysis from those of Rogoff (1985). Specifically, Rogoff finds that some flexibility in monetary policy is invariably better than none (i.e., a constant-money-growth rule or, in terms of their analyses, an extremely perverse monetary authority with $\mu = \infty$).[20] In contrast, the present analysis finds that a constant-growth rule ($N^* = 1$) can emerge as the optimal policy for a wide range of parameter values. There are two reasons for this distinction. First, the present analysis explicitly considers the monetary authority's forecasting ability or lack thereof. Reducing the monetary authority's ability to predict money demand shocks, which should be accommodated to stabilize inflation and output, weakens the case for maintained flexibility in policy. Second, the present analysis allows for the possibility of extending the one-period game setting to a multiperiod game setting where, through the targeting constraint, the monetary authority can partially influence future expectations through its current policy. Even if there were supply shocks, as in Rogoff's model, this distinction would emerge. Nevertheless, for those parameter values under which the present analysis finds that $N^* = \infty$, there could be room for improvement by limiting flexibility through the alternative institutional arrangements studied by Rogoff (1985) and Lohmann (1992): specifically, through the appointment of a perverse central banker with some degree of independence. More generally, as noted in Section III, it might be possible to combine these institutional arrangements with the multiperiod targeting procedure to effect a more efficient balance between credibility and flexibility.

It is difficult to interpret the simulation results directly in terms of actual data because the difference in output goals is measured in (log) levels, and the scale that would be appropriate is not clear. Interpreting the simulation results is more difficult without knowledge of the value of the parameter f. Nevertheless, Table 2 indicates that the full-discretionary solution dominates any finite targeting procedure for a large part of the parameter space.[21] However, for those parameter values in which there is room for improvement by imposing a binding constraint on monetary policy, the table illustrates that, often, $N^* = 1$. This result is rather interesting, considering the fact that, in many nations where central banks have adopted some form of monetary targeting procedure, the targeting horizon coincides with the wage contract period, typically one year.

The result is also illustrated by Figure 2, which depicts regions for $N^* = 1, 2, 3$, and 4 and for $N^* \geq 5$ in terms of combinations of f and y^*/σ_d. Even with this limited opportunity set, the figure shows that the regions for which $N^* = 1$ and $N^* \geq 5$ are the largest areas among those under consideration. More generally, by sequentially increasing the opportunity set and deriving explicit regions for $N^* \geq 5$ (e.g., $N^* = 5$, $N^* = 6$, etc.), it is possible to see that regions for N^* become progressively narrower as $N^* < \infty$ increases. Hence, assuming that N^* is finite, the optimal degree of flexibility is likely to be extremely limited if not entirely eliminated. Of course, this claim implicitly assumes a uniform distribution for the actual

[20]Similarly, Lohmann (1992) finds that complete independence is not optimal (see footnote 19). It should be noted, however, that both Rogoff (1985) and Lohmann (1992) also find that some limits on flexibility are always better than none (i.e., $\mu = 0$ is never optimal). (Lohmann also finds that some independence is always better than none.) This distinction emerges since, in their analyses, limited flexibility ($\mu > 0$) does not constrain reactions to money demand disturbances.

[21]Additional simulations confirm this interpretation of the table. Moreover, they show that, for $f \geq y^*/\sigma_d$, the full-discretionary solution dominates all finite targeting procedures with $N \leq 1{,}000$.

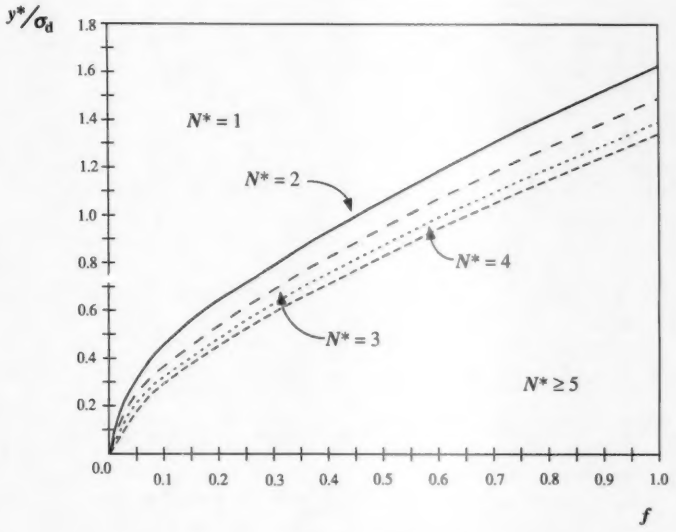


FIGURE 2. OPTIMAL TARGETING HORIZONS

values of the parameters. Without knowledge of the empirical distribution of the key parameters, the figure shows only the conditions (in terms of combinations of parameter values) under which certain regimes would be predicted to emerge by the present analysis. In any case, the simulation exercise provides additional support to the notion that it is important to take into account the credibility problem, in addition to the monetary authority's forecasting ability, when designing *feasible* optimal rules for monetary policy.

One important caveat to the above analysis should be noted. In particular, the assumption that d_t is not serially correlated is a strong assumption. Indeed, the extreme results above (i.e., given N^* is finite, extremely limited flexibility is often desirable) might be driven partly by the transitory nature of the monetary authority's private information. While the effect of persistence of shocks to money demand on the optimal targeting horizon is not obvious in the context of this model, the analysis by Cukierman and Meltzer (1986) suggests that the introduction of such persistence is likely to increase the monetary authority's desire for maintained flexibility. In particular, their analysis shows how persistence in the monetary authority's private information, which is never directly observed by the public, provides a channel through which the monetary authority can influence (indirectly and imperfectly) future expectations, without imposing any explicit limits on flexibility. Persistence, in the context of this model, then might imply a smaller role for a restrictive targeting procedure as an imperfect mechanism by which the monetary authority can influence future expectations indirectly.

However, the qualitative results obtained above are not sensitive to the assumption that the monetary authority does not discount the future. That is, N^* is increasing in f and decreasing in y^*/σ_d, even when the monetary authority discounts the future. In the case of discounting, under a given N-period targeting constraint, the monetary authority's willingness to act on its incentive to create surprise inflation would be greater, and its optimal policy would involve more

complete accommodations of the predicted part of the money demand disturbances, since future reversals are perceived as being less of a burden. As such, when the monetary authority discounts the future, a given N-period targeting solution retains more flexibility but does not gain as much credibility over the full-discretionary solution, relative to when the monetary authority does not discount the future. An examination of the utilities obtained under various N-period targeting constraints when the monetary authority discounts the future suggests that, as the discount rate increases, the optimal targeting horizon, N^*, falls for relatively higher values of y^*/σ_d and smaller values of f; but, for relatively smaller values of y^*/σ_d and larger values of f, an increase in the discount rate will imply an increase in N^*.

III. Concluding Remarks

This paper has investigated the efficacy of average monetary targeting to reduce the inflationary bias that otherwise emerges as a result of the monetary authority's incentive to surprise agents in an effort to increase output beyond its natural level. Reputational considerations and the imposition of a contingent rule, as reasonable methods of eliminating the bias while retaining flexibility for output and inflation stabilization, are called into question when the monetary authority has some private information and there is no mechanism to enforce a truthful dissemination of that information. Provided that the inflationary bias is sufficiently large, an average monetary targeting procedure to eliminate the bias will be appealing. As the targeting horizon gets longer, the amount of flexibility permitted by the procedure increases, but at the cost of a higher variance of the unconditionally expected inflation rate around its socially determined optimal rate. Hence, the optimal targeting horizon defines the optimal trade-off between the cost of limited flexibility and the cost of the remaining inefficiencies of the credibility problem.

As indicated earlier, that the optimal targeting horizon is often very short might be accounted for, in part, by the transitory nature of the monetary authority's private information. Thus an interesting extension of the analysis, left for future research, involves giving that information some persistence and examining the degree to which such persistence will increase the monetary authority's desire for maintaining some degree of flexibility. Also, given this persistence, the analysis could consider the possibility that the monetary authority might employ Crawford and Sobel's (1982) concept of *cheap talk* to reveal its private information, at least partially. Specifically, the monetary authority could make an imprecise announcement (in terms of a range) that partially reveals the persistent (permanent) component of its private information. Accordingly, the monetary authority could keep the average equilibrium inflationary bias low in absolute terms throughout the horizon and, at the same time, enhance the scope of flexibility permitted by the targeting procedure. While the optimal targeting horizon, N^*, would be determined from the stationary part of the model, the degree of precision of the announcements (in terms of the number and width of the ranges) would be determined on the basis of the persistent component of the monetary authority's private information.

Another interesting extension left for future research involves the introduction of (verifiable) supply shocks as in Rogoff (1985), Flood and Isard (1989), and Lohmann (1992) into the model. The richer model specification in a multiperiod setting would permit a fruitful investigation of the merits of the multiperiod targeting procedure relative to alternative institutional solutions—in particular, Rogoff's perverse policymaker solution, Flood and Isard's mixed-strategy solution involving a simple constant-money-growth rule with an *escape clause* to permit temporary (i.e., one-period) reversions to the full-discretionary policy upon the realization of a *seldom experienced* shock, and a combination of those two solutions as suggested by Lohmann. Moreover, the analysis could determine the extent to which these alternative institutional arrangements could be supplemented by the

multiperiod targeting procedure to achieve a more efficient balance between flexibility and credibility.

APPENDIX

This appendix sketches a few steps of the method used to solve the monetary authority's optimization problem subject to the N-period average target constraint. (For simplicity, the tilde ("$\sim$") notation is suppressed here.) In period $t = N$, the monetary authority has no choice. Its policy is determined by (11) in the main text, given the past sequence $\{g_t\}_{t=1}^{N-1}$. In period $t = N-1$, it solves the following problem, recognizing the impact its policy in the current period $t = N-1$ has on future expectations, and subject to (11) in the main text:

$$\text{(A1)}\quad \max_{g_{N-1}} E_{N-1}\left\{ \sum_{t=N-1}^{N} \left[-f(g_t - \delta_t - \pi^*)^2 - (g_t - g_t^e - \delta_t - y^*)^2 \right] \right\}.$$

In light of the constraint on its policy in the next period ($t = N$), the monetary authority does not expect that it can create surprise inflation in the last period. Since (11) implies $g_N^e = g_N$, the last term for $t = N$ in (A1) is irrelevant in the optimization problem. The expectation of the first-order condition, conditional on d_{N-1} is given by

$$\text{(A2)}\quad -2f(g_{N-1} - d_{N-1} - \pi^*) - 2(g_{N-1} - g_{N-1}^e - d_{N-1} - y^*) + 2f\left(N\pi^* - \sum_{t=1}^{N-1} g_t - \pi^* \right) = 0$$

where (11) has been used. Taking the wage-setters' expectation of (A2) and rearranging yields

$$\text{(A3)}\quad g_{N-1}^e = \left(y^*/f + N\pi^* - \sum_{t=1}^{N-2} g_t \right) \Big/ 2.$$

The solution to the monetary authority's problem in period $t = N-1$ is obtained by substituting (A3) back into (A2):

$$\text{(A4)}\quad g_{N-1} = g_{N-1}^e + \frac{1+f}{1+2f} d_{N-1}$$

where g_{N-1}^e is given by (A3).

In period $t = N-2$, the monetary authority solves a similar problem:

$$\max_{g_{N-2}} E_{N-2}\left\{ \sum_{t=N-2}^{N} \left[-f(g_t - \delta_t - \pi^*)^2 - (g_t - g_t^e - \delta_t - y^*)^2 \right] \right\}$$

subject to (11). Under the assumption that the monetary authority recognizes the impact that its current policy has on future expectations, the second terms for $t = N$ and $t = N-1$ in the optimization problem for period $t = N-2$ are irrelevant. Because the monetary authority expects to implement the policy in (A4) in the next period ($t = N-1$) and believes that the wage-setters' expectations of its future ($t = N-1$) policy are consistent with (A4), it recognizes that it cannot create inflation surprises then or any time in the future with its current policy. Using (A3), (A4), and (11) and rearranging, the problem becomes equation (A5), below. Since $E_{N-2}\{d_s\} = 0$ and

$$\text{(A5)}\quad \max_{g_{N-2}} E_{N-2}\left\{ -f(g_{N-2} - \delta_{N-2} - \pi^*)^2 - (g_{N-2} - g_{N-2}^e - \delta_{N-2} - y^*)^2 - f\left\{ \left(\left[y^*/f + (N-2)\pi^* + \sum_{t=1}^{N-2} g_t \right] \Big/ 2 \right) + \frac{1+f}{1+2f} d_{N-1} - \delta_{N-1} \right\}^2 - f\left\{ \left(-\left[y^*/f - (N-2)\pi^* - \sum_{t=1}^{N-2} g_t \right] \Big/ 2 \right) - \frac{1+f}{1+2f} d_{N-1} - \delta_N \right\}^2 \right).$$

$E_{N-2}\{\delta_s\}=0$ for $s>N-2$, the first-order condition to this problem implies

$$(A6)\quad g^e_{N-2}=\left(N\pi^*+2y^*/f-\sum_{t=1}^{N-3}g_t\right)\Big/3$$

and

$$(A7)\quad g_{N-2}=g^e_{N-2}+\frac{2(1+f)}{2+3f}d_{N-2}.$$

Similarly, the optimization problem for period $t=N-3$ has the following solution:

$$(A8)\quad g^e_{N-3}=\left(N\pi^*+3y^*/f-\sum_{t=1}^{N-4}g_t\right)\Big/4$$

$$(A9)\quad g_{N-3}=g^e_{N-3}+\frac{3(1+f)}{3+4f}d_{N-3}.$$

Repeating this exercise for periods $t=N-\rho$, one can verify that the general solution to this problem is given by

$$(A10)\quad g^e_{N-\rho}=\left(N\pi^*+\rho y^*/f-\sum_{\tau=1}^{N-\rho-1}g_\tau\right)\Big/(\rho+1)$$

and

$$(A11)\quad g_{N-\rho}=g^e_{N-\rho}+\frac{\rho(1+f)}{\rho+(\rho+1)f}d_{N-\rho}$$

for $\rho=1,2,3,\ldots,N-1$. The solution given in the text, equation (12), follows from (A10) and (A11), noting that $\rho=N-t$. [A proof via mathematic induction that (12) is the solution is available from the authors upon request.]

REFERENCES

Barro, Robert J. and Gordon, David, "Rules, Discretion and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics*, July 1983, *12*, 101–22.

Canzoneri, Matthew B., "Monetary Policy Games and the Role of Private Information," *American Economic Review*, December 1985, *75*, 1056–70.

Crawford, Vincent and Sobel, Joel, "Strategic Information Transmission," *Econometrica*, September 1982, *50*, 1431–51.

Cukierman, Alex, *Central Bank Strategy, Credibility, and Independence: Theory and Evidence*, Cambridge, MA: MIT Press, 1992.

_____ **and Meltzer, Allan H.,** "A Theory of Ambiguity, Credibility, and Inflation under Discretion and Asymmetric Information," *Econometrica*, September 1986, *54*, 1099–1128.

Flood, Robert and Isard, Peter, "Monetary Policy Strategies," *IMF Staff Papers*, September 1989, *36*, 612–32.

Green, Edward J. and Porter, Robert H., "Noncooperative Collusion Under Imperfect Price Information," *Econometrica*, January 1984, *52*, 87–100.

Hillier, Brian and Malcomson, James, "Dynamic Inconsistency, Rational Expectations, and Optimal Government Policy," *Econometrica*, November 1984, *52*, 1437–51.

Kydland, Finn and Prescott, Edward, "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

Leijonhufvud, Axel, "Constitutional Constraints on the Monetary Powers of Government," in Richard McKenzie, ed., *Constitutional Economics*, Lexington, MA: Lexington Books, 1984, pp. 99–107.

Lohmann, Susanne, "Optimal Commitment in Monetary Policy: Credibility versus Flexibility," *American Economic Review*, March 1992, *82*, 273–80.

Meltzer, Allan H., "The Fed at Seventy-Five," in Michael T. Belongia, ed., *Monetary Policy on the Fed's 75th Anniversary: Proceedings of the Fourteenth Annual Economic Policy Conference of the Federal Reserve Bank of St. Louis*, Boston MA: Kluwer, 1990, pp. 3–65.

Oh, Seonghwan and Garfinkel, Michelle R., "Strategic Considerations in Monetary Policy with Private Information: Can Secrecy Be Avoided?" *Federal Reserve Bank of St. Louis Review*, July/August 1990, *72*, 3–16.

Persson, Torsten and Tabellini, Guido, *Macro-*

economic Policy, Credibility, and Politics, New York: Harwood, 1990.

Rogoff, Kenneth, "The Optimal Degree of Commitment to an Intermediate Monetary Target," *Quarterly Journal of Economics*, November 1985, *100*, 1169–89.

_____, "Reputational Constraints on Monetary Policy," *Carnegie-Rochester Series on Public Policy*, Spring 1987, *26*, 141–82.

Rudin, Walter, *Principles of Mathematical Analysis*, 3rd Ed. New York: McGraw Hill, 1976.

Stein, Jeremy C., "Cheap Talk and the Fed: A Theory of Imprecise Policy Announcements," *American Economic Review*, March 1989, *79*, 32–42.

Why Capital Hires Labor: A Bargaining Perspective

By GREGORY K. DOW*

Organizational forms often serve as vehicles for the appropriation of quasi-rent. Capitalist firms typically emerge when production requires noncontractible investments in specific physical assets, because worker control would divert quasi-rents away from asset owners ex post. Conversely, labor-managed firms tend to emerge in niches requiring specialized human capital but general-purpose physical assets. A key result is that capitalist firms can persist in competitive markets even when labor-managed firms would yield a larger total surplus. (*JEL* C78, D23, J54, P13)

Several decades have passed since Paul Samuelson's (1957) famous remark that in a competitive economy it does not matter whether capital hires labor or the reverse. This claim is surely correct if capital and labor are regarded as nothing more than abstract inputs in a production function. But in reality, the assignment of residual claimant status to capital or labor is not a matter of indifference. Capital-managed firms (KMF's) are much more common than labor-managed firms (LMF's) in Western economies. Except for occasional employee buy-outs of capitalist firms, LMF's tend to cluster in professional, craft, or service activities where little specialized physical capital is required (Raymond Russell, 1985; Avner Ben-Ner, 1988; Henry Hansmann, 1990).

Many different reasons have been suggested for the marginal status of the LMF. The early literature emphasized comparative-static perversities resulting from the (alleged) maximization of net revenue per worker (Benjamin Ward, 1958; Evsey Domar, 1966; Jaroslav Vanek, 1970). Other writers have stressed problems of suboptimal investment deriving from the inability of current LMF members to capture future profits (Eirik Furubotn, 1976; Michael Jensen and William Meckling, 1979). These defects can all be remedied, at least in principle, by making LMF membership rights marketable. This institutional design motivates LMF members to maximize profit or present value rather than income per worker. LMF's thus become isomorphic to capitalist firms in all behavioral respects (Murat Sertel, 1982, 1991; Dow, 1986, 1992, 1993a; Ernst Fehr, 1993). In light of this analytic result, the rarity of LMF's in market economies remains puzzling.[1]

The comparison between capitalist and labor-managed firms is a special but paradigmatic case of a broader question: given free entry, why do some organizational forms survive in equilibrium while others do not? One plausible hypothesis is that surviving modes of organization yield a larger total surplus than potential rivals and, therefore, prove more attractive to factor

*Department of Economics, University of Alberta, Edmonton, Alberta, Canada, T6G 2H4. I am grateful to Gil Skillman, Louis Putterman, Sam Bowles, Herb Gintis, and Margaret Duncan for many stimulating discussions. Marcel Boyer, Christophe Deissenberg, Rick Szostak, and two anonymous referees provided helpful advice on earlier drafts. Financial assistance was supplied by the Social Sciences and Humanities Research Council of Canada.

[1]Other explanations for the dominance of the capitalist firm appeal to informational asymmetries. The classic reference on this subject is Armen Alchian and Harold Demsetz (1972); for a critical review see Louis Putterman (1984). Recent authors have particularly stressed the role of moral hazard between lenders and borrowers (Henry Hansmann, 1988; Mukesh Eswaran and Ashok Kotwal, 1989; Herbert Gintis, 1989).

suppliers. This is essentially the explanation offered by transaction-cost economics (Oliver Williamson, 1985).

In fact, there is no necessary relationship between the size of the surplus generated by an organizational form and its long-run viability. To see why, suppose that production requires an initial investment in a specialized asset, and assume that this investment is noncontractible (perhaps it is too costly to write down all technologically relevant characteristics of the asset). Now consider two possible organizational forms, A and B. Firms of type A maximize aggregate surplus once the specialized asset is in place but make it difficult or impossible for owners of the asset to capture quasi-rent *ex post*. Firms of type B yield a smaller aggregate surplus but direct most of this surplus to the asset's owners. Since type-A firms are unattractive to potential investors, organizers of such firms are disadvantaged relative to type-B rivals. This liability would disappear if the prospective quasi-rent recipients in type-A firms could bribe investors to participate. However, by assumption, one cannot write contracts under which the asset is supplied in exchange for side payments. Given this constraint, the ease with which asset owners can appropriate quasi-rent in a type-B firm proves decisive. Other things equal, firms of this type will prevail in equilibrium because they offer higher returns to investors.[2]

The application to labor-managed firms is straightforward. Suppose that the specific asset is a factory and workers cannot acquire a factory directly because collective capital ownership is impractical. LMF members could nevertheless attempt to lease a factory from external owners. In their managerial capacity the LMF workers would unilaterally choose output levels, working conditions, and the intensity of equipment use. These decision rights could enable workers to transfer substantial quasi-rent to themselves, at the expense of the outside owners. Investors who foresee this problem will be reluctant to build a factory in the first place. One possible solution is for the LMF to extend *ex ante* bribes to potential investors in exchange for construction of the factory. However, workers might not be able to write down all desired characteristics of the factory, or the courts might be unable to verify whether contractual specifications were met. An alternative solution is to let the factory owners manage the firm themselves and capture quasi-rents as residual claimants. While transferring decision rights from workers to owners might reduce total surplus, the factory would at least be built.[3]

Parallel appropriability dilemmas crop up whenever specific investments are noncontractible and the firm's decision-making structure affects the distribution of *ex post* quasi-rent. Due to the noncontractibility problem, all investors must be willing to participate even in the absence of side payments. Organizations in which this *ex ante* participation constraint is violated will not flourish in the long run, regardless of their potential *ex post* productivity (Dow, 1993b).

In this paper I will sketch out one mechanism through which the locus of authority within the firm could influence quasi-rent distribution and thus the viability of alternative organizational forms. The preceding argument about appropriability problems is quite general and does not depend upon the empirical validity of the particular mechanism discussed here, which involves bilateral bargaining between capitalists and workers. The bargaining model is of some interest, however, because it identifies circumstances under which equilibrium organizations fail to maximize aggregate wealth. It is also consistent with the observed distribution of KMF's and LMF's across industries.

[2]A similar logic operates in models in which *ex post* diversion of quasi-rent to unionized workers depresses investment in the capitalist firm (Kim Clark, 1984; Paul Grout, 1984; John Addison and Barry Hirsch, 1989). This effect would vanish if workers and owners could contract on investment levels.

[3]After the factory has been built, workers might try to bribe its owners to adopt the LMF structure if this would increase aggregate surplus. However, once the firm is reorganized as an LMF the external asset owners will have little incentive to maintain the firm's capital stock (see Section V).

I. Bargaining and Organizational Form

The model has the following structure. In the KMF, owners of physical assets are the residual claimants, hire workers, and choose output levels. These roles are reversed in the LMF. Once specialized assets are in place, the hired input bargains over its own remuneration in each organizational form (the wage in the KMF, the rental fee paid to asset owners in the LMF). At the end of this bargaining procedure, the residual claimant unilaterally selects an output. The output chosen depends on the compensation paid to the hired input through a participation constraint.

I then show that there are equilibrium wage bargains in which the KMF does not maximize total surplus, but owners and workers both obtain positive quasi-rent shares. The corresponding LMF bargaining process induces surplus maximization at the production stage, but all of the resulting quasi-rent is appropriated by the firm's work force. The rest of the argument proceeds as before. If physical assets are specialized and noncontractible, the KMF prevails because it allows capital suppliers to recoup their sunk setup costs, even though total surplus is greater in the LMF. The converse also holds: when physical assets are general-purpose but it is impossible to shift the cost of investments in specific human capital to employers, LMF's can prevail in equilibrium even though they fail to maximize aggregate surplus.

Explanations of the capitalist firm which rely on the notion of bargaining power are frequently regarded as analytically loose and faintly disreputable (Williamson, 1980; David Landes, 1986; Eswaran and Kotwal, 1989). A brief defense of this approach is therefore in order.[4] Modern bargaining models rigorously derive the relative power of agents from their rates of time preference or attitudes toward risk (Martin Osborne and Ariel Rubinstein, 1990). For example, workers might place less weight on future income than capitalists because they are rationed in the credit market (Joseph Stiglitz and Andrew Weiss, 1981). If so, they will be handicapped in wage-bargaining. Similarly, workers who cannot diversify their wealth holdings might be more risk-averse than capitalist investors and hence more fearful of a breakdown in negotiations (Jan Svejnar, 1986). Any such reluctance to bear risk will also undermine the bargaining position of workers.

Transaction-cost economics and the bargaining approach share an emphasis upon quasi-rent appropriation. However, they operate at different levels of analysis. In the view of transaction-cost theorists, organizational forms economize on total production and transaction costs (Williamson, 1985). Consequently, the central optimization problem is formulated at the social rather than the individual level. The noncooperative bargaining perspective, by contrast, rejects axioms of collective rationality in favor of a commitment to individual rationality (Rubinstein, 1987). In this approach the firm is a nexus of strategies rather than a nexus of contracts. Organizational forms are shaped by strategic behavior on the part of input suppliers, not just an overarching objective of wealth-maximization (Dow, 1987, 1993b; Williamson, 1987).

Apart from the assumption of noncontractible investment, I will need to introduce some further contracting frictions in the course of the bargaining analysis.

(a) *Output is noncontractible.*—Without this assumption, factor suppliers would always choose output to maximize surplus and then bargain over the division of the resulting quasi-rent. For similar reasons, it is necessary to rule out contracts in which all output-related costs borne by hired inputs are transferred to the residual claimant. I therefore assume that payments to hired inputs are invariant with respect to output (e.g., no piece

[4] For related applications of bargaining theory to questions of organizational form, see Dow (1985, 1988, 1989).

rates). This restriction is appropriate when the courts cannot verify output quality, for example. As will be explained in Section II, the key analytic results also go through when the residual claimant exercises authority over a noncontractible variable other than output, provided that suppliers of the hired input incur an associated cost which cannot be shifted to the residual claimant.

(b) *Hired inputs cannot promise to forgo bargaining in exchange for side payments.*—Section II will show that total surplus varies with the wage. This effect would be of no interest, however, if input suppliers could simply adopt the wage that maximizes overall surplus and then divide the resulting pie through side payments. I exclude this scenario by assuming that if owners bribe workers to accept such deals, workers can simply pocket the bribe and restart negotiations without penalty. Under these conditions it is impossible for agents to avoid the allocative consequences of wage-bargaining. By the same token, I will assume that LMF's cannot bribe owners of physical assets to refrain from opportunistic bargaining over the fee charged for the use of their assets.

(c) *Capital and labor supply cannot be integrated.*—Bargaining inefficiencies between buyers and sellers are often avoided through the device of vertical integration (Benjamin Klein et al., 1978; Williamson, 1985; Sanford Grossman and Oliver Hart, 1986; Hart and John Moore, 1990). Similar inefficiencies could be avoided here either by having capitalists supply labor to their own firms (Eswaran and Kotwal, 1989) or by having the firm's work force collectively acquire its physical assets (Avner Ben-Ner and Byoung Jun, 1990). In what follows, I assume that the first of these options is excluded by the scale of production or the owner's demand for leisure. Worker takeovers are ruled out by liquidity constraints, informational asymmetries, or problems associated with collective capital ownership.[5] As a result, capital and labor are supplied by distinct agents.

Sections II and III develop the bargaining analysis for the KMF. Section II derives the optimal KMF output at an exogenously given wage, assuming that asset owners choose output subject to a worker-participation constraint. Section III endogenizes the wage using a Rubinstein alternating-offers bargaining framework. With reasonable restrictions on the bargaining strategies, I show that two perfect equilibria can arise. The first is the standard Rubinstein solution, in which quasi-rent is divided according to the time preferences of the factor suppliers. Output is excessive in this equilibrium because the firm's owners fail to internalize effort costs borne by workers. In the other equilibrium, output is optimal, workers receive their reservation payoffs, and all quasi-rent flows to the owners of the firm. A dominance result establishes that the Rubinstein solution should be selected when both equilibria exist.

Section IV exploits the symmetry of the KMF and LMF in order to characterize market equilibrium. Assuming that investments in specialized assets are noncontractible, residual claimant status usually goes to the suppliers of such inputs. Equilibrium modes of organization may maximize total surplus; but equilibrium can also involve excessive output by each firm, too little entry, and a diversion of quasi-rent to hired inputs. In particular, KMF's can prevail even when LMF's would generate a larger total surplus. Section V discusses some contractual devices that might promote the emergence of wealth-maximizing organizational forms.

[5] I show elsewhere (Dow, 1993a) that LMF's with collective capital ownership can generate a larger static and dynamic surplus than KMF's, while nevertheless growing more slowly and providing a lower private value per unit of capital.

II. Output Decisions in the Capitalist Firm

Call the factor suppliers in the capitalist firm K and L. These labels denote coalitions of individual agents: for example, shareholders and a labor union. I abstract from issues of input substitution by assuming a Leontief production technology. Labor is supplied to the firm with infinite elasticity *ex ante*, and the reservation utility of workers is normalized at zero. Once K has recruited L, however, some combination of hiring and firing costs leads to a condition of bilateral monopoly (Robert Solow, 1985; Assar Lindbeck and Dennis Snower, 1988).[6]

At the *bargaining* stage, K and L negotiate the fixed wage L will receive when production occurs. K then chooses an output unilaterally at the *production* stage, subject to the constraint that L's payoff be nonnegative.[7] K's payoff as the residual claimant is revenue minus depreciation on the firm's capital stock and the wage paid to L. L's payoff as a hired input is the wage minus the disutility of effort. Overall surplus depends on K's choice at the production stage because depreciation costs and the disutility of effort vary with the output of the firm.

Since K does not directly internalize effort costs, there is an incentive for overproduction. If this tendency is restrained at all, it is only by L's participation constraint. When the wage is high enough, the participation constraint is irrelevant because L has a positive payoff even at K's preferred output. Firm output is consequently excessive. At lower wage levels, the constraint binds, and output is smaller. Through a suitable wage choice, K can be induced to adopt the surplus-maximizing output level. However, L's payoff must be zero if surplus is maximized because the labor-participation constraint always binds in this situation.

The use of output as the residual claimant's choice variable is suggestive, but not essential to the analysis. The key point is that the residual claimant exercises managerial authority over some noncontractible variable (perhaps production technique, or working conditions) which affects costs borne by the hired input. If these costs can only be internalized by the residual claimant through a binding participation constraint, then maximization of total surplus necessarily requires that the hired input receive no quasi-rent. The bargaining analysis of Section III applies with minor modifications to any model of this general type.

The utility functions at the production stage are

$$(1) \qquad u_K(q,w) = pq - c_K(q) - w$$

$$u_L(q,w) = w - c_L(q)$$

where $q \geq 0$ is output and $w \geq 0$ is the wage. The output price p is parametric for an individual firm but will be endogenized when market equilibrium is discussed in Section IV. The function $c_K(\cdot)$ captures depreciation on physical assets, while $c_L(\cdot)$ is the disutility of work effort.

The sum of these payoffs is given by the *surplus function*

$$(2) \qquad s(q) \equiv pq - c_K(q) - c_L(q)$$

$$\equiv u_K(q,w) + u_L(q,w).$$

This is the total quasi-rent available to K and L. In a world of complete contracts the firm would choose output to maximize total surplus, treating $c_K(q) + c_L(q)$ as its cost function. I depart from this framework by assuming that the costs c_K and c_L must be borne privately by the respective input suppliers. A capitalist firm cannot arrange contracts under which the firm's owners bear the disutility of effort, and a labor-managed firm cannot sign contracts with external asset owners under which its membership bears depreciation costs.

[6] Even when L supplies only unskilled labor, K is effectively locked into a bilateral monopoly position if it is costly to replace L *ex post*, and K's existing bargaining position is precisely replicated after a replacement work force has been recruited (Gil Skillman and Harl Ryder, 1988).

[7] If output were chosen prior to (or jointly with) the wage, the wage would merely divide a fixed total surplus between K and L. The resulting wage bargain would depend upon output, an implicit violation of assumption (a) in Section I.

The cost functions are twice continuously differentiable and satisfy the following technical requirements:

ASSUMPTION 1:

$$c_K(0) = c_L(0) = 0$$

ASSUMPTION 2:

$$c'_K(0) = c'_L(0) = 0$$

ASSUMPTION 3:

$$c'_i(q) > 0,\ c''_i(q) > 0 \qquad \text{all } q > 0 \quad i = K, L$$

ASSUMPTION 4:

$$\lim_{q \to \infty} c'_K(q) = \lim_{q \to \infty} c'_L(q) = +\infty.$$

The surplus function is shown in Figure 1. Surplus is strictly concave in output, rising from zero to a maximum at q^* and then returning to zero at q_0. At larger outputs, surplus is negative.

Taking the wage as given at the production stage, K chooses output unilaterally to solve

$$\max_q \{pq - c_K(q) - w\}$$

subject to $c_L(q) \le w$. At high wage levels, the participation constraint does not bind. The resulting unconstrained optimum will be denoted by $\hat{q}_K$. This output level is excessive because K ignores effort costs. At low wage levels, $\hat{q}_K$ is infeasible, and the constraint binds. Writing K's output choice as $q_K(w)$, one therefore has

$$(3)\quad q_K(w) = c_L^{-1}(w) \qquad \text{for } 0 \le w < \hat{w}$$

$$q_K(w) = \hat{q}_K \qquad \text{for } \hat{w} \le w$$

where $\hat{w} \equiv c_L(\hat{q}_K)$ is the lowest wage rate at which L is willing to produce the output $\hat{q}_K$. Since c_L is monotonic, wage increases on the interval $[0, \hat{w}]$ relax the participation constraint and raise output. This effect vanishes once $\hat{q}_K$ becomes feasible.

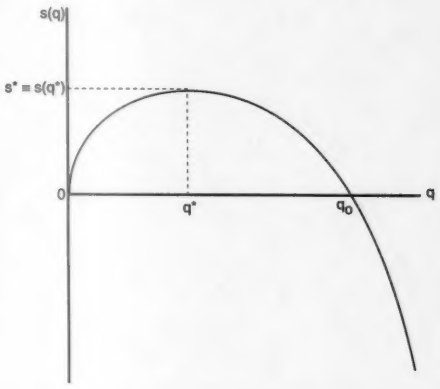


FIGURE 1. SURPLUS AS A FUNCTION OF OUTPUT

Substituting (3) back into (1), one obtains indirect utility functions which depend solely on the wage:

$$(4)\quad v_K(w) \equiv u_K[q_K(w), w]$$

$$v_L(w) \equiv u_L[q_K(w), w].$$

There are two cases to consider, depending on whether or not the participation constraint is binding:

$$(5)\quad \text{if } 0 \le w < \hat{w}, \text{ then } \quad v_K(w) = s[q_K(w)]$$

$$v_L(w) = 0$$

$$\text{if } \hat{w} \le w, \text{ then } \quad v_K(w) = s_K + \hat{w} - w$$

$$v_L(w) = w - \hat{w}$$

where I adopt the notational shorthand $s_K \equiv s(\hat{q}_K)$.

The functions v_K and v_L are shown in Figure 2. In each part of the figure, I distinguish two different situations. To the left of $\hat{w}$, L's payoff is zero since the participation constraint binds. K therefore appropriates the full surplus $s[q_K(w)]$. This surplus attains a maximum of s^* at the wage w^* which induces K to choose the optimal output q^*. Further wage increases will raise

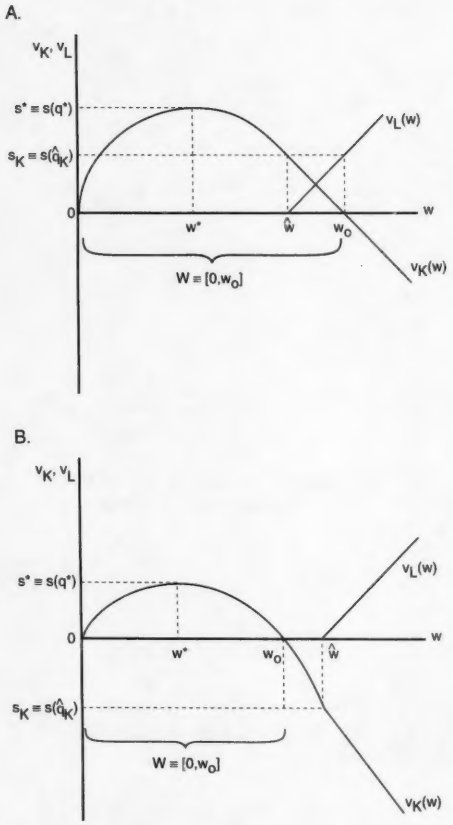


FIGURE 2. INDIRECT UTILITY FUNCTIONS FOR THE KMF A) WHEN $s_K > 0$ AND B) WHEN $s_K < 0$

output and reduce surplus until $\hat{w}$ is reached.

To the right of $\hat{w}$, the participation constraint for labor is irrelevant. K's preferred output choice $\hat{q}_K$ can be implemented, and total surplus is constant at s_K. Variations in the wage in this region serve only to redistribute the fixed surplus between K and L. To the right of w_0, K's payoff becomes negative.

Figures 2A and 2B differ with respect to the sign of the surplus s_K. When K has only a mild incentive to overproduce ($\hat{q}_K < q_0$), the surplus s_K is positive as in Figure 2A. In this case, there are wage bargains in which K and L both have strictly positive payoffs. L has nothing to lose by insisting on such a wage, because the alternative is a payoff of zero. However, any wage high enough to give L a positive payoff necessarily induces K to choose the excessive output $\hat{q}_K$.

If K's incentive to overproduce is stronger ($\hat{q}_K > q_0$), the surplus s_K is negative as in Figure 2B. In this case the labor-participation constraint binds throughout the relevant interval $[0, w_0]$, and L cannot obtain a positive payoff at the bargaining stage. Since L is indifferent and K would like to maximize total surplus, it is plausible to expect an agreement on the wage w^*. With this bargaining outcome, the firm adopts the correct output q^*, and K appropriates the entire (maximized) surplus s^*.

III. Wage Bargaining in the Capitalist Firm

I now formalize these ideas about bargaining outcomes. It is evident from Figure 2 that K and L would like to adopt the wage w^* and then negotiate over the distribution of the surplus s^*. However, this is infeasible by assumption (b) in Section I. We therefore need an explicit model of the procedure through which wages are determined.

I will initially assume $s_K > 0$ as in Figure 2A, since this situation raises more complex bargaining issues. (The case $s_K \leq 0$ is discussed toward the end of this section.) When $s_K > 0$, the set of utility pairs generated by wage offers in the interval $W \equiv [0, w_0]$ is nonconvex, implying that the Nash bargaining solution cannot be used. I turn instead to an alternating-offers game of the Rubinstein (1982) type, where K opens with some proposal $w \in W$ which L either accepts or rejects. If L rejects this offer, L makes a counteroffer in the following period, which K accepts or rejects, and so on. If this process ends with an acceptance, the accepted wage becomes binding, and production occurs. Otherwise each agent receives a zero payoff.

The analysis is straightforward if wage offers are confined to the interval $[\hat{w}, w_0]$ in Figure 2A, since then a pie of fixed size is

being divided. It is well known that the Rubinstein game has a unique perfect equilibrium in this case. However, when offers throughout the interval $W \equiv [0, w_0]$ are admissible, total surplus varies, and neither utility function is strictly monotonic in the wage. This environment generates many perfect equilibria, some of which give L a zero payoff.

There is nevertheless a strong intuition that wage bargains should lie in the range $(\hat{w}, w_0)$ where both payoffs are positive. Any bargaining strategy in which L accepts a zero payoff (e.g., the wage w^*) is dominated by an otherwise identical strategy where L rejects such proposals and insists instead on a positive payoff. If L does hold out for a positive payoff, K is better off with an agreement in the range $(\hat{w}, w_0)$ than with perpetual deadlock.

We formalize this intuition in the following way. Let the one-period discount factors associated with the Rubinstein game be $\delta_K \in (0,1)$ and $\delta_L \in (0,1)$. In addition to the conventional requirement of perfection, we impose two further restrictions on the bargaining strategies of the players.

ASSUMPTION 5: *The equilibrium bargaining strategies are stationary.*[8]

ASSUMPTION 6: *In cases of indifference, existing proposals are accepted.*

Assumption 5 means that each agent's actions are independent of events in earlier bargaining periods. A stationary strategy $\sigma_K = (w_K, z_K)$ is characterized by a wage offer w_K which K always proposes, and a set $Z_K \subseteq W$ of wage offers K is always prepared to accept. A stationary strategy for L is written $\sigma_L = (w_L, z_L)$. The simplicity of such strategies provides an appealing basis for narrowing the set of possible equilibria. Assumption 6 ensures that an agent accepts a proposal when there is no hope of doing any better by prolonging the bargaining process.[9]

PROPOSITION 1: *The only subgame-perfect equilibria in which Assumptions 5 and 6 both hold are the following:*

(i) Extended Rubinstein Solution (ERS): w_K, w_L, Z_K, *and* Z_L *satisfy*

$$w_K = \frac{(1-\delta_L)\hat{w} + \delta_L(1-\delta_K)w_0}{1-\delta_K\delta_L}$$

$$w_L = \frac{\delta_K(1-\delta_L)\hat{w} + (1-\delta_K)w_0}{1-\delta_K\delta_L}$$

$$Z_K = \{w \in W: v_K(w) \geq \delta_K v_K(w_K)\}$$

$$Z_L = [w_K, w_0].$$

(ii) Residual Claimant Optimum (RCO): w_K, w_L, Z_K, *and* Z_L *satisfy*

$$w_K = w^*$$

w_L *is an arbitrary element of* $[0, w_0]$

$$Z_K = \{w \in W: v_K(w) \geq \delta_K s^*\}$$

$$Z_L = [0, w_0].$$

The RCO strategies form a perfect equilibrium if (and only if) $\delta_K s^* \geq s_K$.

(The proof is available from the author upon request.)

The ERS equilibrium is identical to the usual Rubinstein solution, but it extends this solution by permitting any wage offer in the full interval $W \equiv [0, w_0]$ rather than just the truncated set $[\hat{w}, w_0]$ where total surplus is constant. The ERS wage is w_K if K makes the first offer and w_L otherwise. The key

[8]Raquel Fernandez and Jacob Glazer (1991) establish that when a union has the option of striking, some perfect equilibria are inefficient. However, these Pareto-dominated bargaining outcomes involve nonstationary strategies.

[9]These conditions do not by themselves rule out a zero payoff for L. Such outcomes occur in the RCO equilibrium described by Proposition 1. To establish that L's payoff is positive, it is necessary to impose the further condition that the equilibrium strategies are undominated. Proposition 2 can then be used to eliminate equilibria in which L receives a zero payoff.

feature of this equilibrium is that w_K and w_L both lie in the interior of the truncated set $[\hat{w}, w_0]$ even though this restriction was not imposed a priori. Because K fails to internalize the effort cost borne by L, output is excessive ($\hat{q}_K > q^*$), and total surplus falls short of the achievable maximum ($s_K < s^*$).

In an equilibrium of the RCO type, L accepts any wage offer from K. Accordingly, K demands the wage w^* and only accepts wage proposals that are at least as attractive as the maximum surplus s^* received with a one-period delay. L's own wage proposal w_L is arbitrary because L has no reason to make acceptable offers: any wage acceptable to K has zero utility for L.

I turn next to the problem of choosing between the ERS and RCO outcomes. A dominance result indicates that the ERS solution should be chosen and thereby ratifies the earlier intuition that the wage agreement should give both sides a positive payoff.

Definition: The bargaining strategy σ_K is *dominated* if there exists some other strategy σ'_K such that

(a) against any strategy σ_L K's payoff on every subgame from the use of σ'_K is at least as large as the payoff from σ_K; and
(b) there exists some σ_L and some subgame such that K's payoff from σ'_K is greater than the payoff from σ_K.

Otherwise, σ_K is *undominated*.

The strategies in this definition need not be stationary. All subscripts are reversed to define dominance in the case of L.

PROPOSITION 2: *The strategies of K and L in the ERS equilibrium are undominated, but L's strategy in any RCO equilibrium is dominated.*

(The proof is available from the author upon request.) Since bargainers can be expected to discard dominated strategies, this result shows that the ERS equilibrium should be selected.

Assuming that K makes the first wage proposal,[10] one can write the ERS wage bargain more compactly as

$$w_K \equiv \alpha\hat{w} + (1-\alpha)w_0 \tag{6}$$

where $\alpha \equiv (1-\delta_L)/(1-\delta_K\delta_L)$. The power parameter α reflects the time preferences of the two parties.[11] As usually happens in Rubinstein games, more patient agents enjoy a bargaining advantage (w_K is decreasing in δ_K and increasing in δ_L). Upon substituting the wage bargain from (6) back into (5), one obtains the equilibrium KMF payoffs:

$$\text{(7a)} \quad \text{if } s_K > 0, \text{ then } \quad v_K^{KMF} = \alpha s_K$$

$$v_L^{KMF} = (1-\alpha)s_K.$$

Finally, I return to the situation where K's overproduction incentive is strong ($s_K \leq 0$) and Figure 2B applies. Here the bargaining problem disappears, because L's payoff is identically zero on the interval $[0, w_0]$. Intuition suggests the outcome w^*. More formally, since no ERS equilibrium exists for the case shown in Figure 2B, Proposition 2 is irrelevant, and any equilibrium must be of the RCO variety. Proposition 1 indicates that as the time between wage offers approaches zero (δ_K approaches unity), K's acceptance set Z_K shrinks to w^*. If L makes proposals that are acceptable to K when L is indifferent, in the limit w^* is the unique equilibrium bargaining outcome.

[10] The asymmetry introduced by designating K as the first mover vanishes when the time between bargaining proposals goes to zero because w_K and w_L approach a common limit. Write $\delta_K = e^{-\rho_K \Delta}$ and $\delta_L = e^{-\rho_L \Delta}$ where ρ_K and ρ_L are continuous-time discount rates and Δ is the time between proposals. Then,

$$\lim_{\Delta \to 0} w_K(\Delta) = \lim_{\Delta \to 0} w_L(\Delta) = \alpha\hat{w} + (1-\alpha)w_0$$

where $\alpha \equiv \rho_L/(\rho_K + \rho_L)$.

[11] The parameter α could also reflect the risk attitudes of K and L (Ken Binmore et al., 1986; Osborne and Rubinstein, 1990 Ch. 4).

It follows that

$$\text{(7b)} \qquad \text{if } s_K \leq 0, \text{ then } \quad v_K^{KMF} = s^* \qquad v_L^{KMF} = 0.$$

In this case output is optimal, and K appropriates all quasi-rent.

There is a sharp discontinuity in the bargaining outcome at the transition from $s_K > 0$ to $s_K \leq 0$: wages drop abruptly from $\hat{w}$ to w^*, output drops from K's preferred level $\hat{q}_K$ to the surplus-maximizing level q^*, and total surplus jumps from s_K to s^*. The reason is that a nonconvex utility possibility set is replaced by a convex one at $s_K = 0$. Indeed, the set of feasible utility pairs is degenerate for $s_K \leq 0$ since only K can obtain a positive payoff in this case.

I now pause to summarize the main implications derived from the bargaining analysis. When workers cannot obtain a positive payoff under any wage agreement ($s_K \leq 0$), K imposes a wage that supports the optimal output level q^* (through the RCO bargaining equilibrium). If there are some wage agreements that would give workers positive quasi-rent ($s_K > 0$), then such an agreement is actually adopted (through the ERS bargaining equilibrium). In the latter case, the labor-participation constraint is irrelevant, and K chooses the excessive output level $\hat{q}_K$.

One empirical check on the model follows from the prediction that workers should appropriate a share of the capitalist firm's quasi-rent whenever output is excessive. Econometric evidence indicates that workers often do capture substantial rents. This finding can of course be explained in a number of different ways. Efficiency-wage theorists view labor rents as a profit-maximizing response by employers to problems of turnover, adverse selection, or costly supervision (Lawrence Katz, 1986; William Dickens and Katz, 1987; Alan Krueger and Lawrence Summers, 1988). Others have argued that workers capture part of the excess profit resulting from product-market imperfections (Thomas Pugel, 1980; Clark, 1984; Michael Salinger, 1984). While one cannot say with confidence that bargaining processes are responsible, such evidence for the existence of labor rents is at least compatible with the bargaining model developed here.

IV. Equilibrium Organizational Forms

I can now return to the question posed in the Introduction: what determines the equilibrium mode of organization? First, a brief discussion of the LMF is needed. In firms of this type, labor is the residual claimant, and capital is the hired input. Due to this reversal in roles, the payoff functions from (1) are replaced by

$$\text{(1')} \qquad u_K(q,r) = r - c_K(q)$$

$$u_L(q,r) = pq - c_L(q) - r$$

where r is the rental fee paid by L for the use of K's physical assets. Since K and L cannot contract directly on the intensity with which assets are used, r is a lump-sum transfer which does not vary with output. This restriction on LMF's is motivated by Klein et al. (1978), Jensen and Meckling (1979), and Putterman (1988), who emphasize the cost to absentee owners of monitoring the use or abuse of physical assets.

Due to the symmetry of the KMF and LMF, the analysis from Sections II and III goes through for the LMF with an interchange of subscripts. Since the surplus function $s(q)$ defined in (2) does not depend on organizational form, the surplus-maximizing output q^* remains the same. The unconstrained capitalist output ($\hat{q}_K$) is replaced for the LMF by the unconstrained worker output ($\hat{q}_L$) which maximizes the payoff function u_L in (1').

Let $s_L \equiv s(\hat{q}_L)$ be the surplus generated by labor's preferred output. The equilibrium payoffs v_K^{LMF} and v_L^{LMF} from bargaining over the capitalist's rental fee are symmetric to those listed in (7) for the KMF. For easy reference, I state the payoffs from each organizational form in Table 1. To highlight the fact that output choices depend on market prices, I henceforth denote the surplus-maximizing output, the uncon-

TABLE 1—GROSS PAYOFFS TO FACTOR SUPPLIERS FROM THE KMF AND LMF

	KMF	
	$s_K(p) > 0$	$s_K(p) \leq 0$
$v_K^{KMF}(p)$:	$\alpha s_K(p)$	$s^*(p)$
$v_L^{KMF}(p)$:	$(1-\alpha)s_K(p)$	0
	LMF	
	$s_L(p) > 0$	$s_L(p) \leq 0$
$v_K^{LMF}(p)$:	$\alpha s_L(p)$	0
$v_L^{LMF}(p)$:	$(1-\alpha)s_L(p)$	$s^*(p)$

strained KMF output, and the unconstrained LMF output by $q^*(p)$, $\hat{q}_K(p)$, and $\hat{q}_L(p)$. The surplus levels corresponding to these outputs are represented by $s^*(p)$, $s_K(p)$, and $s_L(p)$ in Table 1.[12]

In order for quasi-rents to arise in equilibrium, at least one factor supplier must incur a sunk setup cost prior to the bargaining stage. These costs will be denoted by $y_K \geq 0$ and $y_L \geq 0$.[13] For instance, y_K could be the sunk cost associated with a factory, while y_L could be an expenditure on specialized worker training. I assume that these expenditures cannot be transferred from one factor supplier to another: using the terminology of the Introduction, such investments are noncontractible.[14] This assumption will be discussed at greater length in Section V.

A viable organizational form must give both factor suppliers nonnegative *ex ante* payoffs, so that entrepreneurs who establish such firms can attract all required inputs.

Definition: The KMF is *viable* at a price p if both factor suppliers have nonnegative net payoffs at this price; that is, $v_K^{KMF}(p) \geq y_K$ and $v_L^{KMF}(p) \geq y_L$. LMF viability is defined analogously.

Viability means that all *ex ante* participation constraints are satisfied simultaneously at the going output price. This idea motivates the following definition of market equilibrium.

Definition: Let p_K be the lowest price at which the KMF is viable, and let p_L be the lowest price at which the LMF is viable. Then the KMF is the *equilibrium organizational form* if $p_K < p_L$. Conversely, the LMF is the equilibrium form if $p_L < p_K$. If $p_K = p_L$, then the KMF and LMF can coexist in equilibrium.

Entry by viable organizational forms places downward pressure on price. The preceding definition defines equilibrium organizational forms to be those which survive this competitive pressure in the long run.

Before describing market equilibrium in more detail, I add a further requirement to Assumptions 1–4 from Section II.

ASSUMPTION 7: *The functions $s_K(p)$ and $s_L(p)$ are strictly monotonic.*

Since $s_K(0) = s_L(0) = 0$, Assumption 7 implies that the functions s_K and s_L never change sign. I can therefore refer to the signs of s_K and s_L unambiguously without having to specify the particular output price at which these signs are determined.[15] This facilitates an application of the bargaining results derived in Section III.

[12] If the time between bargaining offers is short, the identity of the first mover is immaterial, and the power parameter α is identical in the KMF and LMF (see footnote 10).

[13] I ignore the case $y_K = y_L = 0$, since then free entry drives price to zero in the product market.

[14] For simplicity I assume that specialized assets are lumpy. The qualitative features of market equilibrium would be the same if investments were modeled continuously, since hired inputs never get positive returns when surplus is maximized. Surplus maximization is thus incompatible with any positive investment on the part of hired inputs (see the remarks on Table 2 in the text). This conclusion is peculiar to the bargaining analysis of Section III. In a different modeling framework, each input might capture positive quasi-rent in each type of firm. The quantitative size of the investments elicited by the KMF and LMF would then need to be examined (an illustrative approach is that of Hart and Moore [1990]).

[15] Assumption 7 holds whenever the cost functions c_K and c_L are quadratic (other special features of the quadratic case are noted at the end of Section IV). The maximum surplus $s^*(p)$ is always strictly increasing by the envelope theorem.

TABLE 2—MARKET EQUILIBRIUM

Case	Equilibrium organization	Output per firm	Equilibrium price	Optimal entry?
A. $s_K < 0$, $s_L < 0$				
1. $y_K > 0$, $y_L = 0$	KMF	$q^*(p_K)$	$s^*(p_K) = y_K$	yes
2. $y_K = 0$, $y_L > 0$	LMF	$q^*(p_L)$	$s^*(p_L) = y_L$	yes
3. $y_K > 0$, $y_L > 0$	none	—	—	—
B. $s_K > 0$, $s_L < 0$				
1. $y_K = 0$	LMF	$q^*(p_L)$	$s^*(p_L) = y_L$	yes
2. $y_K > 0$	KMF	$\hat{q}_K(p_K)$	$s_K(p_K) = \bar{s}$	no
C. $s_K < 0$, $s_L > 0$				
1. $y_L = 0$	KMF	$q^*(p_K)$	$s^*(p_K) = y_K$	yes
2. $y_L > 0$	LMF	$\hat{q}_L(p_L)$	$s_L(p_L) = \bar{s}$	no
D. $s_K > 0$, $s_L > 0$				
1. $p_K < p_L$	KMF[a]	$\hat{q}_K(p_K)$	$s_K(p_K) = \bar{s}$	no
2. $p_K > p_L$	LMF[b]	$\hat{q}_L(p_L)$	$s_L(p_L) = \bar{s}$	no

Notes: Outputs denoted by q^* maximize total surplus at the equilibrium price. Outputs denoted by $\hat{q}$ indicate overproduction. For each set of conditions, the equilibrium price satisfies the given equation.

The surplus level $\bar{s}$ is defined by $\bar{s} \equiv \max\{y_K/\alpha,\ y_L/(1-\alpha)\}$. In the "optimal entry?" column, "yes" means that aggregate surplus cannot be increased through further entry. In such cases, no input supplier receives a rent. A "no" means that entry is suboptimal if and only if $y_K/\alpha \neq y_L/(1-\alpha)$. If $y_K/\alpha > y_L/(1-\alpha)$, then L receives a rent; if this inequality is reversed, then K receives a rent.

[a]When the KMF is an equilibrium in this case, $s_K(p_K) > s_L(p_K)$ so that the KMF surplus is larger at the equilibrium price p_K.

[b]When the LMF is an equilibrium in this case, $s_L(p_L) > s_K(p_L)$ so that the LMF surplus is larger at the equilibrium price p_L.

The following proposition characterizes the minimum price at which each organizational form is viable.

PROPOSITION 3: *Suppose that $i = K$, L is the residual claimant and that Assumptions 1–4 and Assumption 7 hold. The lowest price p_i for which the iMF is viable is determined as follows.*

(a) Suppose $s_i < 0$. Let j be the hired input. Then (i) when $y_j > 0$, the iMF is never viable; but (ii) when $y_j = 0$, the price p_i satisfies $s^*(p_i) = y_i$.

(b) Suppose $s_i > 0$. Then p_i satisfies

$$s_i(p_i) = \bar{s} \equiv \max\left\{\frac{y_K}{\alpha}, \frac{y_L}{1-\alpha}\right\}.$$

PROOF:

(a) From Table 1, the residual claimant's net payoff is $s^*(p) - y_i$, and the hired input's net payoff is $-y_j$. If $y_j > 0$, the iMF is nonviable because the hired input's net payoff is always negative. However, $y_j = 0$ implies that the hired input's net payoff is always zero. By the monotonicity of $s^*(p)$, the lowest price p_i that gives the residual claimant a nonnegative payoff is determined by $s^*(p_i) = y_i$.

(b) From Table 1, K's net payoff in the iMF is $\alpha s_i(p) - y_K$, and L's net payoff in the iMF is $(1-\alpha)s_i(p) - y_L$. Since $s_i(p)$ is monotonic, the lowest price at which both net payoffs are nonnegative is given by the equation stated in part (b).

Table 2 characterizes the market equilibrium corresponding to each sign combination for s_K and s_L. Since the results follow directly from Proposition 3, formal proofs are omitted. Several generalizations emerge from Table 2.

(a) When $s_i < 0$, assigning residual claims and managerial rights to input i leads to

surplus-maximizing output choices. This follows from the nature of the RCO bargaining equilibrium in Section III. Total surplus is maximized by both KMF's and LMF's in case A, LMF's in case B, and KMF's in case C.

(b) Market equilibrium assigns residual claims to specialized inputs (those with positive setup costs) in cases A–C. This is not always compatible with surplus maximization. In cases B2 and C2, surplus maximization demands that hired inputs invest in specialized assets. However, since hired inputs capture no quasi-rent *ex post*, such firms are nonviable.

(c) Entry is optimal when neither input has an equilibrium rent. This implies $s^* - y_K - y_L = 0$ so that net surplus cannot be increased through additional entry. If one input obtains a rent, gross surplus per firm ($\bar{s}$ in the notation of Table 2) exceeds $y_K + y_L$ in equilibrium. Hence, net surplus would be enhanced by further entry. Entry ceases prematurely because the return on investments in specialized assets has already fallen to zero for the input lacking an equilibrium rent.

To see the implications for capitalist and labor-managed firms more clearly, suppose that capital is specialized ($y_K > 0$) but labor is not ($y_L = 0$). In cases A–C of Table 2, the KMF is the equilibrium organizational form (case D is ambiguous). In cases A and C, the KMF also has desirable welfare properties: output maximizes surplus in each firm, entry decisions maximize surplus net of setup costs, and neither input receives a rent. However, in case B, output per firm is too large, entry is suboptimal, and workers capture some quasi-rent. From the standpoint of surplus maximization, the LMF is preferable in case B, but it does not arise because the suppliers of specific physical assets cannot capture any quasi-rent as hired inputs.

By symmetry, identical remarks apply to the LMF when labor is specialized but physical capital is not. In this situation, socially undesirable outcomes occur in case C of Table 2. Here the LMF prevails in equilibrium, but surplus would be larger under capitalist management. KMF's are not viable because they fail to induce investment in specialized human capital.

I close this section with a comment on an important special case of Assumption 7. Suppose the cost functions c_K and c_L have the quadratic form

$$c_K(q) = \beta q^2$$

$$c_L(q) = \gamma q^2.$$

It is easy to check that $\beta > \gamma$ implies $s_K(p) > 0 > s_L(p)$ for all positive prices. When $\beta < \gamma$ these inequalities are reversed. In the former situation case B of Table 2 applies, while in the latter situation case C applies. Since the ambiguous case D is excluded, residual claimant status always goes to suppliers of specialized inputs when costs are quadratic.

V. Wealth Maximization and Organizational Form

Section IV showed that KMF's can prevail in equilibrium even when LMF's yield a larger surplus. To this extent, the bargaining model vindicates proponents of the labor-managed firm, who argue that LMF's may be desirable despite their current marginal status in market economies. However, the comfort provided to LMF advocates is tempered by the fact that a symmetric argument applies to LMF's themselves: some existing LMF's might yield a larger total surplus if they were reorganized as capitalist firms.

Under the bargaining mechanism examined here, hired inputs generally queue up for employment when equilibrium organizational forms fail to maximize total surplus. However, the fundamental difficulty is not the failure of input markets to clear. To see why, suppose case B2 of Table 2 applies, so that only the LMF is surplus-maximizing. Assuming capital is specialized ($y_K > 0$) while labor is not ($y_L = 0$), the KMF prevails in equilibrium. Labor captures a rent of $(1 - \alpha)s_K$, and workers queue for jobs.

If employers could require workers to bid competitively for jobs, they would be compensated *ex ante* for the rent $(1-\alpha)s_K$ and the labor market would clear. This procedure would also address the problem of suboptimal entry, since capitalist investors now capture the full surplus s_K. Even so, the labor-participation constraint remains nonbinding at the production stage, leading to excessive output and a surplus of s_K rather than s^*.

The underlying difficulty can be formulated in either of two ways. One view is that the KMF does not maximize surplus because output is noncontractible [by assumption (a) in Section I]. However, if one takes the noncontractibility of output as a given, then the problem boils down instead to one of noncontractible investment. Surplus would be maximized if the LMF displaced the KMF, but this does not occur because LMF's do not attract investment in physical capital. Since setup costs are borne noncooperatively, capital suppliers will not undertake specialized investments that bring them no return at the bargaining stage, even if such investments enable workers to appropriate large positive returns. Problems associated with shared returns from noncooperative investment decisions are commonplace in the economics of organization (see e.g., Hart and Moore, 1990). Because hired inputs cannot obtain positive returns when surplus is maximized, here one has a particularly stark version of this phenomenon: if the hired input incurs *any* positive setup cost, firms maximizing total surplus cannot be viable. A discontinuity appears when this setup cost goes to zero, since then it is unnecessary for the hired input to appropriate any quasi-rent. Accordingly, the surplus-maximizing mode of organization becomes viable.[16]

Whether or not proper investment incentives can be restored depends on the available contracting options. One solution is to have LMF members compensate external capital suppliers for their sunk costs. Even if K and L both know the characteristics of the assets put in place by K, however, the courts might find it impossible to verify these characteristics. K's promise to supply assets in exchange for a side payment is worthless if L has no guarantee of compliance. Such contracts may also be too costly or complex for the parties to formulate in the first place.

Of course, once a capitalist firm has been established and specialized assets are already in place, there is an incentive to reassign residual claims to labor if the LMF would yield a larger surplus. However, when capital obtains no quasi-rent as a hired input and investment is noncontractible, there is no incentive for the outside asset owners to maintain the firm's capital stock once it has been reorganized as an LMF. The long-run survival of the LMF is accordingly undermined.[17] In practice, employee takeovers of capitalist firms frequently occur when the firm is already facing financial difficulty. Maintenance of the existing capital stock (whether by outsiders or workers themselves) is seldom a central concern when a takeover involves a failing firm.

An alternative scenario is for K to create any needed assets unilaterally and then sell LMF membership rights to workers. By paying an entry fee to K, workers would obtain residual claims on the firms's profit stream, the right to manage the firm, and the right to negotiate a fee for the use of K's physical assets. If there are no frictions in the market for membership and the LMF is surplus-maximizing, K will extract the surplus $s^*(p)$ *ex ante* and will voluntarily incur

[16]This discontinuity in the market equilibrium does not result from the lumpy character of investments. Rather, it follows from the fact that hired inputs receive zero quasi-rent when surplus is maximized (see footnote 14).

[17]The analogous problem for capitalist firms, that of inducing workers to acquire specialized skills, is frequently solved by having employers bear training costs or through informal training on the job. Because skills can often be maintained simply by using them, incentives for the maintenance of such assets seldom pose a serious dilemma in the capitalist firm.

the sunk cost y_K whenever $s^*(p) \geq y_K$. Entry decisions are optimal, and output is chosen correctly *ex post*. Since workers receive a zero payoff (net of membership fees) in equilibrium, the labor market clears. Markets for LMF membership thus have some appeal (Dow, 1986, 1992, 1993a).

This approach removes the need for third-party verification, since now prospective LMF members can directly inspect the firm's assets before deciding whether to pay the membership fee demanded by K. However, incentives for K to maintain or expand the firm's capital stock remain problematic (unless workers periodically bid to renew their membership rights, a procedure which raises other difficulties if workers gain firm-specific skills after the first round of bidding). The LMF membership market is also vulnerable to informational asymmetries between K and L, since asset owners have incentives to exaggerate the actual value of their assets. Indeed, a similar valuation problem arises for employee buy-outs of existing capitalist firms (Ben-Ner and Jun, 1990).

Finally, LMF's might combine the roles of capital and labor supplier by treating physical assets as the collective property of the enterprise. Under this solution, the LMF no longer hires capital from outside owners, and the bargaining problem vanishes. The main objection concerns the investment incentives of the LMF. Several authors have asserted that LMF's with collective capital ownership will underinvest (Furubotn, 1976; Jensen and Meckling, 1979). In theory, this defect can be cured by making membership rights marketable, but the practicality of this proposal remains uncertain (Dow, 1993a).

Further progress in understanding why capital hires labor thus appears to require a better understanding of LMF membership markets and specifically the nature of the impediments to such markets. This in turn will demand a better integration of the bargaining perspective with standard adverse-selection and moral-hazard concepts. Work along these lines should help to explain the prevalence of capitalist firms and may suggest improvements in the design of their labor-managed counterparts.

REFERENCES

Addison, John and Hirsch, Barry, "Union Effects on Productivity, Profits, and Growth: Has the Long Run Arrived?" *Journal of Labor Economics*, January 1989, *7*, 72–105.

Alchian, Armen and Demsetz, Harold, "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, *62*, 777–95.

Ben-Ner, Avner, "Comparative Empirical Observations on Worker-Owned and Capitalist Firms," *International Journal of Industrial Organization*, March 1988, *6*, 7–31.

_____ **and Jun, Byoung,** "Buy-Out in a Bargaining Game with Asymmetric Information," mimeo, Industrial Relations Center, University of Minnesota, December 1990.

Binmore, Ken, Rubinstein, Ariel and Wolinsky, Asher, "The Nash Bargaining Solution in Economic Modelling," *Rand Journal of Economics*, Summer 1986, *17*, 176–88.

Clark, Kim, "Unionization and Firm Performance: The Impact on Profits, Growth, and Productivity," *American Economic Review*, December 1984, *74*, 893–919.

Dickens, William and Katz, Lawrence, "Inter-industry Wage Differences and Industry Characteristics," in Kevin Lang and Jonathan Leonard, eds., *Unemployment and the Structure of Labor Markets*, Oxford: Blackwell, 1987, pp. 48–89.

Domar, Evsey, "The Soviet Collective Farm as a Producer Cooperative," *American Economic Review*, September 1966, *56*, 743–57.

Dow, Gregory, "Internal Bargaining and Strategic Innovation in the Theory of the Firm," *Journal of Economic Behavior and Organization*, September 1985, *6*, 301–20.

_____, "Control Rights, Competitive Markets, and the Labor Management Debate," *Journal of Comparative Economics*, March 1986, *10*, 48–61.

_____, "The Function of Authority in Transaction Cost Economics," *Journal of Economic Behavior and Organization*, March 1987, *8*, 13–38.

_____, "Information, Production Deci-

sions, and Intra-firm Bargaining," *International Economic Review*, February 1988, *29*, 57–79.

_____, "Knowledge Is Power: Informational Precommitment in the Capitalist Firm," *European Journal of Political Economy*, 1989, *5*, 161–76.

_____, "Markets for Membership in the Labor-Managed Firm: A Simple General Equilibrium Model," mimeo, Department of Economics, University of Alberta, March 1992.

_____, (1993a) "Democracy versus Appropriability: Can Labor-Managed Firms Flourish in a Capitalist World?" in Samuel Bowles, Herbert Gintis, and Bo Gustafsson, eds., *Democracy and Markets: Problems of Participation and Efficiency*, New York: Cambridge University Press, 1993 (forthcoming).

_____, (1993b) "The Appropriability Critique of Transaction Cost Economics," in Christos Pitelis, ed., *Markets, Hierarchies, and Transaction Costs: Critical Assessment*, Oxford: Blackwell, 1993 (forthcoming).

Eswaran, Mukesh and Kotwal, Ashok, "Why Are Capitalists the Bosses?" *Economic Journal*, March 1989, *99*, 162–76.

Fehr, Ernst, "The Simple Analytics of a Membership Market in a Labour-Managed Economy," in Samuel Bowles, Herbert Gintis, and Bo Gustafsson, eds., *Democracy and Markets: Problems of Participation and Efficiency*, New York: Cambridge University Press, 1993 (forthcoming).

Fernandez, Raquel and Glazer, Jacob, "Striking for a Bargain Between Two Completely Informed Agents," *American Economic Review*, March 1991, *81*, 240–52.

Furubotn, Eirik, "The Long-Run Analysis of the Labor-Managed Firm: An Alternative Interpretation," *American Economic Review*, March 1976, *66*, 104–23.

Gintis, Herbert, "Financial Markets and the Political Structure of the Enterprise," *Journal of Economic Behavior and Organization*, May 1989, *11*, 311–22.

Grossman, Sanford and Hart, Oliver, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, August 1986, *94*, 691–719.

Grout, Paul, "Investment and Wages in the Absence of Binding Contracts: A Nash Bargaining Approach," *Econometrica*, March 1984, *52*, 449–60.

Hansmann, Henry, "Ownership of the Firm," *Journal of Law, Economics, and Organization*, Fall 1988, *4*, 267–305.

_____, "When Does Worker Ownership Work? ESOPs, Law Firms, Codetermination, and Economic Democracy," *Yale Law Journal*, June 1990, *99*, 1749–1816.

Hart, Oliver and Moore, John, "Property Rights and the Nature of the Firm," *Journal of Political Economy*, December 1990, *98*, 1119–58.

Jensen, Michael and Meckling, William, "Rights and Production Functions: An Application to Labor-Managed Firms and Codetermination," *Journal of Business*, October 1979, *52*, 469–506.

Katz, Lawrence, "Efficiency Wage Theories: A Partial Evaluation," in Stanley Fischer, ed., *NBER Macroeconomics Annual*, 1986, Cambridge, MA: MIT Press, pp. 235–89.

Klein, Benjamin, Crawford, Robert and Alchian, Armen, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

Krueger, Alan and Summers, Lawrence, "Efficiency Wages and the Inter-industry Wage Structure," *Econometrica*, March 1988, *56*, 259–93.

Landes, David, "What Do Bosses Really Do?" *Journal of Economic History*, September 1986, *46*, 585–623.

Lindbeck, Assar and Snower, Dennis, *The Insider – Outsider Theory of Employment and Unemployment*, Cambridge, MA: MIT Press, 1988.

Osborne, Martin and Rubinstein, Ariel, *Bargaining and Markets*, New York: Academic Press, 1990.

Pugel, Thomas, "Profitability, Concentration and the Inter-industry Variation in Wages," *Review of Economics and Statistics*, May 1980, *62*, 248–53.

Putterman, Louis, "On Some Recent Explanations of Why Capital Hires Labor," *Economic Inquiry*, April 1984, *22*, 171–87.

_____, "The Firm as Association Versus

the Firm as Commodity: Efficiency, Rights, and Ownership," *Economics and Philosophy*, October 1988, *4*, 243–66.

Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, January 1982, *50*, 97–109.

_____, "A Sequential Strategic Theory of Bargaining," in Truman Bewley, ed., *Advances in Economic Theory — Fifth World Congress*, New York: Cambridge University Press, 1987, pp. 197–224.

Russell, Raymond, "Employee Ownership and Internal Governance," *Journal of Economic Behavior and Organization*, September 1985, *6*, 217–41.

Salinger, Michael, "Tobin's *q*, Unionization, and the Concentration–Profits Relationship," *Rand Journal of Economics*, Summer 1984, *15*, 159–70.

Samuelson, Paul, "Wages and Interest: A Modern Dissection of Marxian Economic Models," *American Economic Review*, December 1957, *47*, 884–912.

Sertel, Murat, "A Rehabilitation of the Labor-Managed Firm," in Murat Sertel, ed., *Workers and Incentives*, Amsterdam: North-Holland, 1982, pp. 11–26.

_____, "Workers' Enterprises in Imperfect Competition," *Journal of Comparative Economics*, December 1991, *15*, 698–710.

Skillman, Gil and Ryder, Harl, "Bargaining with Costly Unilateral Replacement," Working Paper No. 88-17, Department of Economics, Brown University, June 1988.

Solow, Robert, "Insiders and Outsiders in Wage Determination," *Scandinavian Journal of Economics*, 1985, *87* (2), 411–28.

Stiglitz, Joseph and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

Svejnar, Jan, "Bargaining Power, Fear of Disagreement, and Wage Settlements: Theory and Evidence from U.S. Industry," *Econometrica*, September 1986, *54*, 1055–78.

Vanek, Jaroslav, *The General Theory of Labor-Managed Market Economies*, Ithaca, NY: Cornell University Press, 1970.

Ward, Benjamin, "The Firm in Illyria: Market Syndicalism," *American Economic Review*, September 1958, *48*, 566–89.

Williamson, Oliver, "The Organization of Work," *Journal of Economic Behavior and Organization*, March 1980, *1*, 5–38.

_____, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

_____, "Transaction Cost Economics: The Comparative Contracting Perspective," *Journal of Economic Behavior and Organization*, December 1987, *8*, 617–25.

Advance-Purchase Discounts and Monopoly Allocation of Capacity

By Ian L. Gale and Thomas J. Holmes*

Optimal pricing by a monopoly airline that faces capacity constraints during the peak demand period is studied. The existence of capacity constraints means that in order to expand output the airline must divert demand from the peak period to the off-peak period. A particular advance-purchase discount policy is shown to be the profit-maximizing method of selling tickets. If the advance-purchase requirement were infeasible, output and total surplus would both be lower. (*JEL* L12)

Substantial discounts for air travel are generally available to travelers who are willing to purchase tickets a week or more in advance of the departure date. This paper demonstrates why the imposition of an advance-purchase requirement may be a profit-maximizing strategy for a monopoly airline facing capacity constraints during times of peak demand.[1] The practice of imposing an advance-purchase requirement on discount tickets is also placed in a favorable light in terms of its impact on social welfare.

Two features of demand are important for the analysis that follows. First, there is demand uncertainty at the individual level. Consumers are not certain in advance of their desired departure time. For example, a business traveler may not know when a business meeting will end. Second, consumers vary in terms of the disutility or "time cost" that they suffer if they take a flight that leaves at the wrong time. For example, business travelers are thought to have a high value of time compared to tourists.

The existence of capacity constraints during peak demand periods means that in order to expand output the monopolist must divert demand from the peak period to the off-peak period, using some form of discount as an inducement. One way to accomplish this diversion is to offer a uniform discount for off-peak flights. Under this scheme, all individuals taking off-peak flights obtain a discount, including business travelers with a high value of time who turn out to prefer an off-peak departure time. An advance-purchase requirement for the off-peak discount keeps such high-time-cost individuals from obtaining the discount, enabling the monopolist to extract greater rents from them. Because the advance-purchase requirement enables the monopolist to target the discount more precisely, there is more of an incentive to expand output by diverting demand to the off-peak period. As a result, total output and total surplus are higher than they would be if the advance-purchase restriction were

*Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706. We thank David Butz, Kenneth Hendricks, Michael Riordan, and Pablo Spiller for valuable discussions. We also acknowledge the many helpful comments and suggestions of two referees. Finally, we are grateful to seminar participants at Boston University, the University of British Columbia, Georgetown University, the University of Illinois, the University of North Carolina, the University of Wisconsin, the Antitrust Division of the Department of Justice, and the Winter 1990 meetings of the Econometric Society.

[1] The idea that restricted discount fares may serve to divert demand from the peak to the off-peak period has been suggested by Steven Salop (1978), Severin Borenstein (1983), and Elizabeth Bailey et al. (1985), all of whom observed that the availability of discounts is often quite limited during peak demand periods. However, this earlier literature does not explain why airlines employ advance-purchase discounts rather than offering a lower, uniform fare for off-peak flights, for example.

infeasible. (The effect on consumer surplus is ambiguous, however.)

We also endogenize the choice of capacity. Increases in peak capacity allow the airline to attain a given level of aggregate output while diverting fewer consumers to the off-peak period. This enhances the firm's ability to extract informational rents from consumers, so a monopoly airline is inclined to set capacity at higher levels than a welfare-maximizing social planner would. However, if the advance-purchase requirement were infeasible, the monopolist would choose a level of capacity that is even more excessive from a social-welfare standpoint.

In many models of imperfect price discrimination a "bad" is bundled with the good in order to facilitate surplus extraction by a monopoly seller. For example, a low price may be available only if the consumer is willing to spend time in a queue or travel to a distant store (see Raymond Chiang and Chester Spatt, 1982). Such schemes can be effective when a consumer's reservation value for the product is positively correlated with her value of time. Moving from uniform pricing to price discrimination may be welfare-enhancing in such cases, but the aggregate time cost borne by consumers represents a deadweight loss to society. In the current paper, the advance-purchase requirement serves to divert demand in the face of a capacity shortage. The diversion of consumers to the off-peak flight facilitates the expansion of output, and the socially optimal way to accomplish this expansion is to divert individuals with low values of time. As a consequence, the aggregate time cost does not represent a deadweight loss to society here.

The basic demand and supply conditions are described in Section I. In Section II we formulate the seller's problem as a mechanism design problem and solve for the optimal direct-revelation mechanism. We then show that the optimal mechanism is implemented by a pricing policy that offers an advance-purchase discount fare for the off-peak flight. Section III examines advance-purchase policies in more detail and shows why they dominate policies in which uniform discounts are given for the off-peak flight. Endogenous determination of capacity is discussed in Section IV.

I. The Model

Two flights operate each day on a particular route. One departs at a time of high demand, the other at a time of low demand. There is a continuum of risk-neutral consumers who each demand one ticket, and the total measure of consumers is equal to 1. Between $t = 0$ and $t = 1$, a measure $\alpha \in (\frac{1}{2}, 1)$ of consumers learn that they will prefer the "high-demand" flight. The remainder learn that they will prefer the "low-demand" flight. (We also refer to the flights as the "peak" and the "off-peak.") Since there is a continuum of consumers, there is no aggregate demand uncertainty. Each flight has capacity K, where $\frac{1}{2} < K < \alpha$. Since $K < \alpha$, there is not enough capacity to serve all consumers who prefer the high-demand flight. However, total capacity is sufficient to serve all consumers. We normalize marginal costs to zero and set fixed costs to zero as well.

Each consumer has a reservation value $r > 0$ for her preferred flight, but consumers differ in terms of the value of their time. A "type-y" consumer suffers a time cost $y \in [0, \bar{y}]$ if she flies at her less-preferred time, which gives her a reservation value of $r - y$ for the less-preferred flight. The distribution of consumer types is the same for those who prefer the peak as for those who prefer the off-peak.[2] The density of types, $f(\cdot)$, is strictly positive on the support $[0, \bar{y}]$. The cumulative distribution is the continuous function $F(\cdot)$.

II. The Optimal Monopoly Policy

There are many ways that an airplane could sell tickets for its flights. One strategy would be to post prices, which could de-

[2] Preferences for the peak flight could also be positively correlated with value of time [i.e., $\alpha'(y) > 0$], which is plausible since business travelers may be more likely to prefer the peak departure time than are tourists. The results are qualitatively similar in this case.

pend on the time of the flight and the time of purchase. The firm could also sell tickets through some form of auction. Given the wide array of possible strategies, a complete analysis of this problem would appear to be a daunting task. However, the literature on monopoly pricing under incomplete information has made great progress in dealing with problems of this type (see Milton Harris and Artur Raviv, 1981; Roger Myerson, 1981; Eric Maskin and John Riley, 1984). The papers in this literature consider a monopolist's choice of a "direct-revelation mechanism" in which consumers report their private information to the seller. All subsequent allocation decisions and payments depend on these reports. The major benefit of this approach is its generality: all members of a broad class of pricing mechanisms, including posted-price schemes and auctions, are equivalent to some direct-revelation mechanism.[3]

This section formulates the mechanism-design problem and solves for the optimal direct-revelation mechanism. It is shown that the optimal mechanism is characterized by the "cutoff" y^m. All individuals with time costs below y^m depart on the low-demand flight. Those individuals with time costs above y^m who prefer the low-demand flight take that flight. Those individuals with time costs above y^m who prefer the high-demand flight take that flight or do not fly. A particular posted-price policy, which offers an advance-purchase discount on the low-demand flight, is then shown to be equivalent to the optimal direct-revelation mechanism.

A. *The Optimal Direct-Revelation Mechanism*

All contracting takes place at $t = 0$. The mechanism that the seller announces must induce consumers to report their private information truthfully. In particular, they must find it optimal to divulge their time cost at $t = 0$ and their flight preference at $t = 1$. It turns out that the latter constraint is not binding at the optimum. We first examine a relaxed problem in which the flight preference of each consumer is observable at $t = 1$. We subsequently show that the solution to the relaxed problem satisfies the truth-telling constraint at $t = 1$, so a solution to the relaxed problem is a solution to the original problem.

The mechanism specifies the probability that an individual departs on a given flight as a function of her reported time cost and her observed flight preference. Let $\theta_j^i(y)$ be the probability that an individual departs on flight j when she reports a time cost y and is observed to prefer flight i. From the consumer's perspective, one need only focus on two derived probabilities. The first is the *ex ante* probability that the consumer departs on the wrong flight:

$$(1)\quad w(y) \equiv (1-\alpha)\theta_{\mathrm{H}}^{\mathrm{L}}(y) + \alpha\theta_{\mathrm{L}}^{\mathrm{H}}(y).$$

The first term on the right-hand side of (1) gives the probability that she prefers the low-demand flight but ends up taking the high-demand flight. [Recall that $1-\alpha$ is the *ex ante* probability of preferring the low-demand flight, while $\theta_{\mathrm{H}}^{\mathrm{L}}(y)$ is the probability of taking the high-demand flight, conditional on preferring the low-demand flight.] The second term covers the opposite case. The other derived probability is the probability of departing on one flight or the other:

$$(2)\quad d(y) \equiv (1-\alpha)\left[\theta_{\mathrm{L}}^{\mathrm{L}}(y) + \theta_{\mathrm{H}}^{\mathrm{L}}(y)\right] + \alpha\left[\theta_{\mathrm{L}}^{\mathrm{H}}(y) + \theta_{\mathrm{H}}^{\mathrm{H}}(y)\right].$$

A consumer's net surplus also depends on the expected payment made to the seller. Let $\pi(y)$ be the expected payment from a consumer who reports a time cost of y.[4]

[3]The "revelation principle" ensures that attention can be restricted to direct-revelation mechanisms in which it is a Bayes-Nash equilibrium to report one's type truthfully. See Myerson (1981) or Drew Fudenberg and Jean Tirole (1991) for more details.

[4]Here, $\pi(y)$ could be a certain payment that is collected at $t = 0$, or it could be the expected value of a random payment based on contingencies at $t = 1$. In the mechanism discussed in Subsection II-B, only those consumers who fly will actually pay.

The expected net surplus of an individual whose true type is y', but who reports y, is

$$(3)\quad V(y,y') \equiv d(y)r - \pi(y) - w(y)y'.$$

Incentive compatibility (IC) requires that it be optimal to report one's time cost truthfully at $t = 0$:

$$y' \in \underset{y \in [0,\bar{y}]}{\operatorname{argmax}} V(y,y').$$

Truthful revelation at $t = 0$ gives an expected net surplus of $v(y) \equiv V(y,y)$. The first lemma states that IC can be expressed in terms of two conditions on $w(\cdot)$ and $v(\cdot)$. In particular, individuals with higher time costs are no more likely to fly at the wrong time than are consumers with lower time costs. (For the method of proof see Myerson [1981] or Fudenberg and Tirole [1991].)

LEMMA 1: *IC holds if and only if $w(\cdot)$ is nonincreasing and $v(\cdot)$ satisfies the following condition:*

$$(4)\quad v(y) = \int_y^{\bar{y}} w(u)\,du + v(\bar{y}).$$

Consumers enter into a contract only if their expected return from doing so is nonnegative. This is the voluntary participation (VP) constraint:

$$v(y) \geq 0 \quad \text{for all } y \in [0,\bar{y}].$$

Since $w(\cdot) \geq 0$, if $v(\bar{y}) \geq 0$ then VP holds whenever (4) is satisfied. The monopolist will extract all surplus from the highest type, so $v(\bar{y}) = 0$ in an optimal mechanism. We henceforth assume that (4) holds and that $v(\bar{y}) = 0$. When $w(\cdot)$ is nonincreasing, IC and VP will then be satisfied.

Using (3) and (4), and the fact that $v(\bar{y}) = 0$, the expected payment from a type-y consumer can be written as

$$(5)\quad \pi(y) = d(y)r - w(y)y - \int_y^{\bar{y}} w(u)\,du.$$

Together, the first two terms represent the net social surplus, while the third term is the net private surplus that accrues to the type-y consumer. Substituting in (5) and integrating by parts gives the total revenue for the seller:

$$(6)\quad \int_0^{\bar{y}} f(y)\pi(y)\,dy = \int_0^{\bar{y}} (f(y)\,d(y)\cdot r - w(y)[yf(y) + F(y)])\,dy.$$

Finally, define the level of output on flight j by

$$(7)\quad Q_j = \int_0^{\bar{y}} f(y)\left[(1-\alpha)\theta_j^{\mathrm{L}}(y) + \alpha\cdot\theta_j^{\mathrm{H}}(y)\right] dy \qquad (j = \mathrm{L},\mathrm{H}).$$

The firm's problem can now be stated formally as expression (8), below, subject to:

(C1) technical constraints: $\theta_j^i(y) \geq 0$ for all i and j and y; $\theta_{\mathrm{L}}^i(y) + \theta_{\mathrm{H}}^i(y) \leq 1$ for all i and y;
(C2) monotonicity: $w(\cdot)$ is nonincreasing;
(C3) resource constraints: $Q_j \leq K (j = \mathrm{L},\mathrm{H})$.

A regularity condition on the distribution of types simplifies the analysis that follows. First, define the function $\rho(\cdot)$ by

$$(9)\quad \rho(y) \equiv r - y - \frac{F(y)}{f(y)}.$$

The regularity condition is the following:

(10) $\rho(\cdot)$ is monotone decreasing.

$$(8)\quad \max_{\theta_{\mathrm{L}}^{\mathrm{L}}(\cdot),\theta_{\mathrm{H}}^{\mathrm{L}}(\cdot),\theta_{\mathrm{L}}^{\mathrm{H}}(\cdot),\theta_{\mathrm{H}}^{\mathrm{H}}(\cdot)} \int_0^{\bar{y}} (f(y)d(y)r - w(y)[y\cdot f(y) + F(y)])\,dy$$

It will be shown in Section III that (10) ultimately requires that marginal revenue, properly defined, be monotonic.

The cutoff y^m is now defined in terms of two critical time costs. First, let y^0 be the unique root of $\rho(y)=0$. [If $\rho(\bar{y})>0$, then $y^0 \equiv \bar{y}$.] Second, let y^K be the unique cutoff that satisfies

$$(11) \qquad \alpha\left[1-F(y^K)\right]=K.$$

There will be a measure K of consumers who prefer the high-demand flight and who have a time cost above y^K. The optimal mechanism is characterized by the smaller of these two values: $y^m \equiv \min\{y^0, y^K\}$. The result is now stated.

LEMMA 2: *Assume that the regularity condition (10) is satisfied. The following probability functions solve the seller's problem:*

$$(12a) \quad \theta_L^L(y)=1 \text{ and } \theta_H^L(y)=0 \quad \text{for all } y$$

$$(12b) \quad \theta_L^H(y)=1 \text{ and } \theta_H^H(y)=0 \quad \text{for } y \le y^m$$

$$(12c) \quad \theta_L^H(y)=0 \text{ and}$$

$$\theta_H^H(y)=\frac{K}{\alpha[1-F(y^m)]} \quad \text{for } y>y^m.$$

The proof is in the Appendix.

When capacity is large, few consumers need to be diverted to the low-demand flight in order to have a total output of $Q=1$, so $y^K \le y^0$. In this case the optimal cutoff is $y^m=y^K$, and there are just enough seats on the peak flight for all types $y>y^m$. That is, $\theta_H^H(y)=1$ for all $y>y^m$. When capacity is small, $y^K>y^0$, which means that $y^m<y^K$. Demand exceeds supply in this case, and seats on the peak flight are rationed. (The allocation among these consumers does not affect anyone's expected payoff. The optimal mechanism described above uses random rationing. There exist other solutions, all of which are payoff-equivalent for all individuals.)

A consumer with a time cost $y>y^m$ flies with an *ex ante* probability $d(y)=(1-\alpha)+\alpha\theta_H^H(y)$. The probability of flying at the wrong time is $w(y)=0$ for all types $y>y^m$, so these consumers get zero expected surplus. The expected payment for a consumer with a time cost $y>y^m$ is therefore

$$(13) \quad \pi(y)=\left[(1-\alpha)+\alpha\theta_H^H(y)\right]r.$$

Consumers with $y \le y^m$ all take the off-peak flight with probability 1. The expected payment for a consumer of type $y \le y^m$ is therefore

$$(14) \quad \pi(y)=r-\alpha y-\alpha(y^m-y)$$

$$=r-\alpha y^m.$$

The optimal direct-revelation mechanism for the relaxed problem in which flight preferences are publicly observable at $t=1$ has been characterized. Individuals would have no incentive to misrepresent their flight preferences in the optimal mechanism (see footnote 5 for further discussion). Hence, the solution above also solves the original problem in which the truth-telling constraint at $t=1$ is imposed.

B. *The Optimality of Advance-Purchase Discounts*

We now describe a posted-price policy which is equivalent to the optimal direct-revelation mechanism described above. The firm sets a "coach" fare $p_C=r$, which is paid by all consumers who purchase at $t=1$ and by anyone who purchases a ticket at $t=0$ for the peak flight. The firm also sets a discount fare $p_D=r-\alpha y^m$ for tickets on the off-peak flight that are purchased at $t=0$. The firm limits the refundability of the discount tickets. Specifically, the refund is less than $r-y^m$. Finally, it is assumed that resale of tickets is impossible.

A consumer facing this fare structure would never purchase a ticket for the peak flight at $t=0$, since she has probability $1-\alpha$ of preferring the off-peak flight. Either she will wait until $t=1$ to purchase, and receive zero surplus, or she will purchase a ticket for the low-demand flight at $t=0$. Since the latter strategy imposes an expected time

cost of αy on a type-y consumer, her expected net surplus from purchasing in advance is

$$(15) \qquad r - p_D - \alpha y = \alpha(y^m - y).$$

All consumers with $y \le y^m$ will purchase a seat on the low-demand flight at $t = 0$. Anyone who turns out to prefer the high-demand flight is better off taking the low-demand flight rather than seeking the partial refund. All consumers with a time cost above y^m will wait until $t = 1$ to purchase and will attempt to purchase a ticket for their preferred flight. The expected payments of all consumers are identical to their payments in the optimal direct-revelation mechanism, and the allocation is the same.[5] We conclude with the following proposition.

PROPOSITION 1: *Assume that the regularity condition (10) is satisfied. The optimal selling mechanism is implemented by a policy in which (i) the price of a ticket purchased at $t = 0$ for the off-peak flight is $r - \alpha y^m$, and (ii) the price of all other tickets is r.*

III. Analysis of the Optimal Policy

This section provides additional intuition underlying the results of Section II. Consider the class of advance-purchase discount (APD) policies that specify a price $p_D \le r$ for a ticket on the off-peak flight that is purchased at $t = 0$, and a price $p_C = r$ for all other tickets. We restrict attention to APD policies here since the results of Section II show that a particular APD policy implements the optimal mechanism.

Suppose that the discount fare is equal to $\hat{p}_D$, and let $\hat{y}$ be the time cost of the marginal consumer who purchases in advance. It follows from (15) that $\hat{p}_D = r - \alpha\hat{y}$. (We can also restrict attention to policies in which $\hat{y} \le y^K$.) It is convenient here to view the monopolist as choosing the cutoff $\hat{y}$. Profit can then be written as

$$(16) \quad \Pi(\hat{y}) = r\{K + (1-\alpha)[1 - F(\hat{y})]\} + (r - \alpha\cdot\hat{y})F(\hat{y}).$$

The two terms account for revenues from the sale of coach tickets and discount tickets, respectively. For $\hat{y} < y^K$, the slope is

$$(17) \quad \Pi'(\hat{y}) = \alpha f(\hat{y})\left[r - \hat{y} - \frac{F(\hat{y})}{f(\hat{y})}\right].$$

Note that $\Pi'(0) > 0$. As $\hat{y}$ increases, aggregate output expands, but the price of a discount ticket falls. When $\hat{y} = 0$, the discount price is $\hat{p}_D = r$, so there is no first-order loss in revenue from existing purchasers when the discount fare is lowered. However, there is a first-order increase in sales, so the profit-maximizing level of advance-purchase sales is strictly greater than zero.

When the cutoff satisfies $\hat{y} \le y^K$, total output is

$$(18) \quad Q(\hat{y}) \equiv (1-\alpha) + K + \alpha\cdot F(\hat{y})$$

which is strictly increasing in $\hat{y}$ for $\hat{y} < y^K$. The cutoff can now be expressed as $\hat{y}(Q)$, and $\hat{y}'(Q) = [\alpha\cdot f(\hat{y}(Q))]^{-1}$. Taken together, equations (17) and (18) indicate that the marginal revenue from increasing output is

$$(19) \quad \frac{d\Pi}{dQ} = r - \hat{y}(Q) - \frac{F(\hat{y}(Q))}{f(\hat{y}(Q))} = \rho(\hat{y}(Q))$$

for $Q < 1$. Hence, the regularity condition requires that marginal revenue be monotone decreasing.[6] The optimum can now be

[5] It is now clear that no consumer has an incentive to misrepresent her flight preference in the optimal direct-revelation mechanism. Individuals with $y > y^m$ have a clear incentive to report their flight preference truthfully, since otherwise they might end up on the wrong flight. Individuals with $y \le y^m$ take the low-demand flight, whatever report they make about their flight preference, so they have no incentive to lie.

[6] When marginal revenue and $\rho(\cdot)$ are not monotonic, lotteries are used to flatten out $\rho(\cdot)$. It is now possible that the monopolist will offer two advance-purchase prices, with quantity rationing at the lower price. For a related result see Jeremy Bulow and John Roberts (1989), for example.

interpreted as follows: output is expanded by selling off-peak advance-purchase discount tickets until marginal revenue is zero or until the boundary is reached and all consumers are buying.

The method of diverting demand from peak periods to off-peak periods that has been most commonly studied in the literature is explicit peak-load pricing. A uniform discount is given to all consumers who take the off-peak flight. We will refer to such a price structure as a uniform, off-peak discount (UOP) policy. If the capacity on the two flights were controlled by a continuum of "atomless" firms rather than a single monopolist, then the competitive equilibrium price structure would entail a uniform, off-peak discount. In the remainder of this section we explain why an advance-purchase discount policy is superior to a uniform, off-peak discount policy for a profit-maximizing monopolist.

Suppose the seller were to offer a UOP policy in which the prices of the peak and off-peak flights were, respectively, r and $r - y^m$.[7] As in the optimal mechanism, demand for off-peak seats consists of (i) consumers who prefer the off-peak flight, and (ii) consumers who prefer the peak but who have time costs below y^m. Letting $\Pi^u(\cdot)$ denote profit as a function of the cutoff when a UOP policy is employed,

$$(20)\quad \Pi^u(y^m) = (r - y^m)[(1-\alpha) + \alpha F(y^m)] + rK$$

which is lower than the profit from the APD policy.

It is instructive to examine the sources of this difference in profits. The revenue from the peak flight is rK in both cases. However, the revenues from the off-peak flight differ for two reasons. First, in the APD regime the discount only goes to the consumers who purchase at $t = 0$. With a uniform, off-peak discount, the discount goes to *all* passengers who take the off-peak flight. Second, the advance-purchase price in the APD regime is $r - \alpha \cdot y^m$, which is strictly greater than $r - y^m$. The marginal consumer who purchases in advance in the APD regime must be compensated for the *expected* disutility associated with taking the off-peak flight. If consumers are induced to switch flights after they know their preferred flight, as in the UOP regime, then the actual disutility of the marginal consumer who switches determines the discount. Hence, for a given level of output, the size of the discount and the measure of consumers who receive it are both smaller under the APD policy.[8]

The differences between these two regimes are also germane when one studies the effect of banning advance-purchase sales. Suppose that the seller were restricted to using policies with a uniform, off-peak discount. When the peak price is equal to r and the off-peak price is equal to $r - \hat{y}$, the change in profit as $\hat{y}$ varies is

$$(21)\quad \Pi^{u\prime}(\hat{y}) = \alpha f(\hat{y})\left[r - \hat{y} - \frac{F(\hat{y})}{f(\hat{y})}\right] - (1-\alpha).$$

For the same cutoff $\hat{y}$, the increase in profit as the cutoff is increased is lower here than when an APD policy is used. The profit-maximizing cutoffs from the optimal mechanism (y^m) and from the UOP regime (y^u) are graphed in Figure 1, as functions of capacity. Output is $Q = 1$, and the cutoff equals y^K when capacity exceeds K^m and K^u in the respective regimes.

The incentive to increase the cutoff is greater in the APD regime for the very reasons given above. Under both an APD policy and a UOP policy, increasing the cutoff by $\Delta\hat{y}$ causes output to rise by $\alpha f(\hat{y})\Delta\hat{y}$. The increase in the discount that is required to attain this rise in output is

[7]It can be shown that, when the monopolist is constrained to offer a UOP price structure, he will set price equal to r for the peak flight.

[8]The optimal mechanism can also be implemented by a policy with uniform prices $r - y^m$ for the off-peak flight and r for the peak, and an "entry fee" of $(1-\alpha)y^m$. The entry fee is paid by *all* individuals, including those who ultimately do not fly, before flight preferences are known.

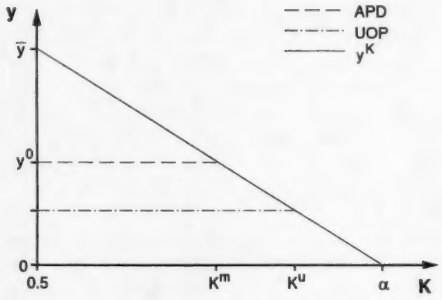


FIGURE 1. PROFIT-MAXIMIZING CUTOFFS AS FUNCTIONS OF CAPACITY

only $\alpha\Delta\hat{y}$ in the APD regime, compared to $\Delta\hat{y}$ in the UOP regime. This increase in the discount also goes to fewer recipients in the APD case. Marginal revenue is lower when the seller is restricted to using policies with a uniform off-peak discount, so output and total surplus are both lower. The effect on consumer welfare is ambiguous: in some cases all consumers are strictly worse off (*ex ante*) with APD; in other cases there is a Pareto improvement with APD.[9]

IV. Endogenous Determination of Capacity

The analysis so far has taken the level of capacity as given. We now endogenize capacity. Suppose that K units of capacity cost $c(K)$, where $c(\cdot)$ is strictly increasing and convex, and $c'(\frac{1}{2}) < r$. Suppose also that the regularity condition (10) is satisfied. There are two main results. First, the level of capacity chosen by a monopolist is at least as great as the level chosen by a social planner. Second, total surplus is higher, and capacity is lower, for a monopolist who is able to offer advance-purchase discounts compared to one who can only offer a uniform, off-peak discount.

As our benchmark, we consider a social planner who selects capacity and the allocation of seats to maximize total surplus. (The results will be qualitatively the same if the planner can only select capacity and the monopolist then selects prices.) When capacity is already in place, and the cutoff is $\hat{y}$, total surplus is

$$(22)\quad \mathrm{TS}(\hat{y}) = [K + (1-\alpha) + \alpha \cdot F(\hat{y})]r - \alpha\int_0^{\hat{y}} f(y)\,y\,dy.$$

The first term represents the gross surplus, while the second represents the aggregate time cost. The planner will divert all consumers with time costs $y \le y^*(K) \equiv \min\{r, y^K\}$ from the peak flight to the off-peak flight. When $r < y^K$, some consumers with time costs $y > r$ who prefer the peak flight will not fly. When $r \ge y^K$, sufficiently many consumers are induced to switch that $Q = 1$.

Now consider the choice of capacity. The maximized value of total surplus, as a function of the level of capacity chosen, can be written as

$$(23)\quad \mathrm{TS}^*(K) = [K + (1-\alpha) + \alpha F(y^*(K))]r - \alpha\int_0^{y^*(K)} f(y)\,y\,dy - c(K).$$

Since $y^*(K) = \min\{r, y^K\}$, the slope is

$$(24)\quad \mathrm{TS}^{*\prime}(K) = \min\{r, y^K\} - c'(K).$$

Note that $\mathrm{TS}^{*\prime\prime}(K) < 0$ for all $K \in (\frac{1}{2}, \alpha)$.

Now consider a monopolist who will use the optimal APD policy. Profit, as a function of the level of capacity chosen, can be

[9] If $K > K^u$ in Figure 1, then the optimal cutoff equals y^K in both regimes, and total surplus is the same in the two regimes. Since the monopoly profit is strictly higher with APD, consumer surplus must then be lower. Consumers of type $y < y^K$ are worse off in the APD regime because the discount for the off-peak flight in the APD regime is smaller than the off-peak discount in the UOP regime. Consumers of type $y > y^K$ are worse off since they receive no surplus in the APD regime whereas in the UOP regime there is a positive probability that they will prefer the off-peak flight and receive a discount. For other parameter values, the APD regime is Pareto-superior to the UOP regime. When r is small, for example, the optimal policy in the UOP regime is to set the off-peak price equal to r, so consumer surplus is zero for all individuals.

written as

$$(25)\quad \Pi(K) = r\{K + (1-\alpha)[1 - F(y^m)]\} + (r - \alpha y^m)F(y^m) - c(K).$$

Straightforward calculations indicate that the slope is

$$(26)\quad \Pi'(K) = \min\left[r,\ y^K + \frac{\alpha - K}{\alpha f(y^K)}\right] - c'(K).$$

Since the monopolist values capacity at least as highly at the margin, the level of capacity chosen by the monopolist will be at least as great as the level chosen by the social planner. (Note, however, that the monopolist does not have excess capacity on the peak flight. The level of capacity chosen under APD is strictly below α, since there is no first-order loss in revenue from increasing the cutoff above zero, and the optimal cutoff is never larger than y^K.)

When r is relatively small, the level of capacity chosen satisfies $r = c'(K)$ for the planner and the monopolist. In both cases, aggregate output is $Q < 1$, so increasing capacity simply leads to an increase in sales for the peak flight. As r rises, the capacities increase, until $r = y^K = c'(K)$. [The value of K at which $y^K = c'(K)$ is clearly unique.] Beyond this point, the monopolist's capacity continues to rise, while the planner's capacity remains constant. Eventually, a point is reached at which the monopolist's capacity itself reaches a plateau. In the region where the planner and the monopolist choose the same capacity, output is strictly greater for the planner. In the region where the planner's capacity is strictly lower, the planner's output ($Q = 1$) is at least as great.

The idea that a monopolist would actually invest in more capacity than would a social planner may seem counterintuitive at first. An explanation for this result is given for the case of large r. When K is large, output is $Q = 1$ for the monopolist and the planner both. Suppose, then, that r is sufficiently large that we are in that region. The benefit to a social planner of increasing capacity at the margin by ΔK is a reduction of $\Delta K \cdot y^K$ in the aggregate time cost, since ΔK fewer passengers need to be diverted to the off-peak flight. Under the APD scheme as well, ΔK fewer passengers need to be diverted, so $(\Delta K/\alpha)$ fewer advance-purchase seats will be sold. These $(\Delta K/\alpha)$ passengers no longer receive the discount $\alpha \cdot y^K$, so this represents a gain to the seller of $\Delta K \cdot y^K$. However, there is an additional gain, since the size of the discount will drop as the time cost of the marginal consumer drops. As a consequence, the incentive to expand capacity is greater for the monopolist.

Now return to the case of a monopolist who is restricted to offering policies with a uniform discount for the off-peak flight. The capacity chosen here will be at least as large as when advance-purchase discounts are permitted. To see this, note that profit in the UOP regime is of the form:

$$(27)\quad \Pi^u(K) = rK + (r - y^u)[(1-\alpha) + \alpha \cdot F(y^u)] - c(K)$$

so the slope is

$$(28)\quad \Pi^{u\prime}(K) = \min\left[r, y^K + \frac{1-K}{\alpha f(y^K)}\right] - c'(K).$$

The increase in profit from adding capacity at the margin is at least as great in the UOP regime as in the APD regime.

From the standpoint of total surplus maximization, the level of capacity chosen in the UOP regime is even more excessive than in the APD regime. The reason for this is clear. There are two ways for a monopolist to expand output. One is to divert more consumers to the off-peak flight; the other is to add more capacity. For a given K, it is relatively less costly to expand output by diverting consumers to the off-peak flight when advance-purchase discounts are used, so there is less of an incentive to expand capacity in the APD regime.

While capacity is lower in the APD regime, total surplus is strictly greater. There are two cases to consider, depending on whether r is small or large. When r is small, capacity is the same in the two regimes, and there is excess demand for seats on the peak flight. Following the analysis of Section III, output and total surplus are both strictly greater in the APD regime when r is small. As r increases, the APD policy reaches an output of $Q=1$ first. (Hence, output is always at least as large in the APD regime as in the UOP regime.) Let $\bar{K}^a$ and $\bar{K}^u$ be the capacities chosen in the two regimes when r is large. Total surplus in the APD regime is equal to $TS^*(\bar{K}^a)$, since the level of output ($Q=1$) is socially optimal given the capacity chosen. Total surplus in the UOP regime is no larger than $TS^*(\bar{K}^u)$. The fact that $TS^*(\cdot)$ is strictly concave and the fact that $K^* < \bar{K}^a < \bar{K}^u$ [where K^* maximizes $TS^*(K)$] ensures that total surplus is strictly greater in the APD regime than in the UOP regime.

V. Concluding Comments

Suppose that it is uncertain initially which flight will be the high-demand flight. The optimal direct-revelation mechanism remains the same as in Proposition 1, but implementation via advance-purchase discounts is slightly different. Now, the flight time for which the advance-purchase ticket is valid is not announced until after the demand uncertainty is resolved. This type of policy, whereby a consumer pays in advance and is subsequently served at a time chosen by the seller, is seen in the Australian airline industry, for example.[10] Other situations in which consumers accept a lower probability of being served at their preferred time involve third-class mail and some electric-power contracts (see Robert Wilson, 1989).

The model still predicts that no seats on the peak flight are sold at a discount. In reality, airlines do make some discount seats available on peak flights. In another paper (Gale and Holmes, 1992), we argue that if the timing of peak demand is uncertain, and if offering contingent tickets of the type just described is infeasible, then the firm will sell advance-purchase discount tickets for both flights. Advance-purchase discounts smooth out demand fluctuations in that setting.

In addition to the advance-purchase requirement for discount tickets, airlines usually require that consumers purchase a round-trip ticket that includes a Saturday-night stayover. Suppose that some high-value-of-time individuals (i.e., business travelers) were certain in advance that their preferred flight was the off-peak. An advance-purchase requirement by itself would not exclude such individuals from receiving the discount. Suppose, however, that an observable characteristic, such as an unwillingness to stay over a Saturday night, were positively correlated with value of time. The firm might wish to impose further restrictions in order to exclude such individuals from receiving a discount.

The main empirical prediction of this paper is that airlines will limit the availability of discount seats on peak flights. There is anecdotal evidence that airlines do this, but more careful study is needed.[11] Ideally, such a study would employ a flight-level data set that specifies, for a given date and time, the fares paid by all of the passengers on that flight. Unfortunately, this is greater detail than is available in the airline data sets currently in use. (An alternative strategy would be to collect data on the availability of discount seats by using the on-line computer reservation systems now available.)

A finding that discount seats are in limited supply during peak times is certainly consistent with the hypothesis that they serve a demand-diversion role. However, if high-reservation-value consumers are more likely to prefer to fly during times of peak demand, then the availability of discounts

[10] We thank Jeff Borland for this example.

[11] Airlines sometimes have stated explicitly that certain advance-purchase discounts are available only on off-peak flights ("Eastern to Offer Discounts on Off-Peak Flights in East" [*Wall Street Journal*, 9 August 1987, p. 10]).

will be limited during peak times, even without capacity constraints. The validity of the demand-diversion hypothesis would require that the flights with limited availability of discount seats also be the flights that tend to be up against capacity constraints.

APPENDIX

PROOF OF LEMMA 2:

First, substitute equations (1) and (2) for $w(\cdot)$ and $d(\cdot)$ into the objective function (8). If we ignore the monotonicity constraint (C2) and the resource constraint (C3) on the low-demand flight, we obtain the (continuous) linear program (LP) in expression (A1), below, subject to

$$\text{(A2)} \quad \theta_L^L(y) + \theta_H^L(y) \le 1 \quad \text{for all } y$$

$$\text{(A3)} \quad \theta_L^H(y) + \theta_H^H(y) \le 1 \quad \text{for all } y$$

$$\text{(A4)} \quad \int_0^{\bar{y}} f(y)[(1-\alpha)\theta_H^L(y) + \alpha\theta_H^H(y)]\,dy \le K$$

$$\text{(A5)} \quad \theta_j^i(y) \ge 0 \quad \text{for all } i, j, \text{ and } y.$$

Let $\lambda^L(y)$ and $\lambda^H(y)$ be the multipliers on constraints (A2) and (A3), and let μ be the multiplier on the resource constraint (A4). The dual of the linear program (LP*) is

$$\text{(A6)} \quad \min_{\lambda^L(\cdot),\lambda^H(\cdot),\mu} \int_0^{\bar{y}} [\lambda^L(y) + \lambda^H(y)]\,dy + K\mu$$

subject to

$$\text{(A7)} \quad \lambda^L(y) \ge f(y)(1-\alpha)r$$

$$\text{(A8)} \quad \lambda^L(y) + f(y)(1-\alpha)\mu \ge f(y)(1-\alpha)\rho(y)$$

$$\text{(A9)} \quad \lambda^H(y) \ge f(y)\alpha\rho(y)$$

$$\text{(A10)} \quad \lambda^H(y) + f(y)\alpha\mu \ge f(y)\alpha r$$

$$\text{(A11)} \quad \lambda^L(y) \ge 0 \quad \lambda^H(y) \ge 0 \quad \text{for all } y \qquad \mu \ge 0.$$

We claim that a solution to the primal problem (LP) is given by (12a)–(12c) and that a solution to the dual problem (LP*) is given by

$$\text{(A12)} \quad \lambda^L(y) = f(y)(1-\alpha)r \quad \text{for all } y$$

$$\text{(A13)} \quad \lambda^H(y) = f(y)\alpha\rho(y) \quad \text{for } y \le y^m$$

$$\lambda^H(y) = f(y)\alpha\rho(y^m) \quad \text{for } y > y^m$$

$$\text{(A14)} \quad \mu = r - \rho(y^m).$$

The proposed solution (12a)–(12c) to LP is clearly feasible. The proposed solution to LP* is also feasible since $\rho(y) \ge 0$ for $y \le y^m$. It is easily verified that the complementary slackness conditions hold. Therefore, (12a)–(12c) solve LP. Since this solution satisfies the monotonicity constraint (C2) and the resource constraint (C3) on the low-demand flight, it solves the maximization problem (8).

REFERENCES

Bailey, Elizabeth E., Graham, David R. and Kaplan, Daniel P., *Deregulating the Airlines*, Cambridge, MA: MIT Press, 1985.

Borenstein, Severin, "Price Discrimination in Free-Entry Markets," Ph.D. dissertation, Massachusetts Institute of Technology, 1983.

Bulow, Jeremy and Roberts, John, "The Simple Economics of Optimal Auctions," *Journal of Political Economy*, October 1989, *97*, 1060–90.

Chiang, Raymond and Spatt, Chester, "Imperfect Price Discrimination and Welfare," *Review of Economic Studies*, April 1982, *49*, 155–81.

Fudenberg, Drew and Tirole, Jean, *Game Theory*, Cambridge, MA: MIT Press, 1991.

Gale, Ian and Holmes, Thomas, "The Efficiency of Advance-Purchase Discounts in the Presence of Aggregate Demand Uncertainty," *International Journal of Indus-*

$$\text{(A1)} \quad \max_{\theta_L^L(\cdot),\theta_H^L(\cdot),\theta_L^H(\cdot),\theta_H^H(\cdot)} \int_0^{\bar{y}} f(y)[(1-\alpha)r\cdot\theta_L^L(y) + (1-\alpha)\rho(y)\theta_H^L(y) + \alpha\rho(y)\theta_L^H(y) + \alpha r\theta_H^H(y)]\,dy$$

trial Organization, 1992, *10* (4), 413–37.

Harris, Milton and Raviv, Artur, "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *American Economic Review*, June 1981, *71*, 347–65.

Maskin, Eric and Riley, John, "Monopoly with Incomplete Information," *Rand Journal of Economics*, Summer 1984, *15*, 171–96.

Myerson, Roger, "Optimal Auction Design," *Mathematics of Operation Research*, February 1981, *6*, 58–73.

Salop, Steven C., *Alternative Reservation Contracts*, Office of Economic Analysis Staff Memorandum, Washington DC: Civil Aeronautics Board, 1978.

Wilson, Robert, "Efficient and Competitive Rationing," *Econometrica*, January 1989, *57*, 1–40.

Auctions With Endogenous Valuations: The Persistence of Monopoly Revisited

By KALA KRISHNA*

It is shown that the standard arguments for a monopoly to persist break down when many units of capacity become available sequentially. The reason is that deterring entry at one stage affects the cost of doing so in later stages. This force can be so powerful that entry is deterred only at the final stage. The persistence of monopoly or its erosion over time depends on the timing of arrival of capacity, the size of early versus subsequent capacity availability, the length of the capacity-acquisition stage relative to the life of the industry, and the discount factor. (*JEL* C72, D42, D44, F13, L12)

When additional scarce resources become available over time, will these be acquired by a large incumbent firm in order to perpetuate its market power, or will these be acquired by new entrants to dissipate that power? If the additional resources become available as a single lump, a standard argument (Richard J. Gilbert and David M. G. Newbery, 1982) suggests that these resources have more value to the incumbent than to a new entrant. Thus the incumbent would outbid the entrant, and monopoly would persist. In this paper, I argue that if the additional resources become available piecemeal rather than as a whole, this reasoning needs to be amended. I study this problem by formulating it as an auction in which successive units of industry-specific capacity are bid for by the firms in the industry.

The model developed here also applies to markets where output is limited by direct quantity controls, such as licenses to operate or sell, on both the entry of new firms and the expansion of existing firms. In many cities, taxicab operation is limited to those who hold medallions. More medallions are created and allocated occasionally. Similarly, takeoff and landing slots limit the number of flights at major airports. New slots are created by adding to airport capacity by building new runways or terminals and are valued differently by different firms. Quotas limit the level of imports directly, and import licenses have been used in a number of countries to implement quotas. Licenses have been allocated directly or sold, as in Colombia and New Zealand. Quota license auctions have also been proposed in the United States as a way of raising revenue (see C. Fred Bergsten et al., 1987; Krishna, 1990a).[1]

The examples given above share some common features. First, the *value* that a firm, incumbent or entrant, attaches to an additional unit of the resource depends on the distribution of the remaining units across firms. For example, how a monopolist with 100 units of existing capacity will value the tenth of an additional ten units depends on whether the first nine units were acquired by the monopolist, acquired by a single rival, or evenly distributed across many firms. Second, since the units are made available sequentially, a firm's *willingness to pay* for

*Fletcher School of Law and Diplomacy, Mugar 250, Tufts University, Medford, MA 02155. I am grateful to Vijay Krishna for all his help, to Richard J. Gilbert for directing me to the paper by Tracy R. Lewis, to a referee for many useful comments, to the National Science Foundation for research support under Grant No. SE8822204, and to the Institute for International Economics in Stockholm for their hospitality in June 1989 when work on this paper was begun.

[1]Robert C. Feenstra et al. (1990) look at the implications of standard auction theory for quota license auctions.

earlier units depends on how it anticipates later units to be allocated. This, of course, is derived from the values placed on the final distributions of capacity. I adopt the term *endogenous valuations* to refer to the first of these aspects, namely, that since the values are determined endogenously from the outcome of market interaction they depend on the allocation of all units.[2]

Endogenous valuations occur naturally when profits are interrelated either through costs or through prices. Price interrelations are to be expected with imperfect competition, as argued above. Cost interrelations can also lead to similar effects since the costs of production can depend on the allocation of the scarce resource. For example, consider the value of the right to drill a well to access a common pool of oil. The value of an additional drilling right depends on the overall allocation of such rights. This is because the pressure of oil in the common pool, and so the resulting cost of extraction, depends on the extraction levels of all parties.

The valuations that result from such considerations take on forms that are different from those often postulated in auction theory. In the literature on multiple-unit auctions, the standard assumptions made are that the marginal valuations of successive units are constant and independent of the allocation of other units. Alternatively, it is assumed that each bidder has a downward-sloping demand curve.[3] Neither formulation seems appropriate when the valuations are determined as a result of market interaction. This, rather than whether the valuations are determined endogenously or specified exogenously, is the key. In order to focus on these issues, in what follows, I assume that there is no uncertainty about the various aspects of the situation and that the agents are completely informed. Here I depart from a major concern of auction theory, namely, the role of incomplete information.

More closely related is some recent work by Ian Gale (1990a), who looks at the auction of multiple objects with superadditive values. He shows that, under certain conditions, revenue is maximized when a single bidder acquires all the objects. Carolyn Pitchik (1989) studies sequential auctions with incomplete information, in which, although valuations are fixed, effective valuations of agents are endogenously determined because of the presence of budget constraints. Effective valuations are like effective demand in that they are valuations backed by purchasing power.

The sequential-auction model is used to address the question of whether there are intrinsic forces at work that will lead to a concentration of scarce resources and ensure that monopolization occurs. The standard argument (Gilbert and Newbery, 1982) relies on the idea that, since competition destroys industry profits, deterring entry is more valuable to an incumbent than entry is to a potential entrant. While the basic idea is simple and powerful, there may be other factors that mitigate this effect. These include: uncertainty about the technology of preemption or the entrant's characteristics; superiority of the entrant's costs or technology; and when the scarce resource is knowledge, the absence of well-defined property rights. If these effects are large, monopoly power may not persist.[4]

However, none of these factors leads to an erosion of monopoly power in this paper. Rather, the model developed here points to another aspect of the problem, namely, that the threat of entry occurs at more than one point in time. The idea is that deterring entry at one stage raises the cost of doing so at future stages, and this force can be so powerful that entry is not deterred except at the final stage!

In an insightful and closely related paper, Tracy R. Lewis (1983) makes a similar argument. However, his analysis overestimates

[2]Such effects have been noted before. For example, Michael L. Katz and Carl Shapiro (1986) recognize that valuations are endogenous in studying how to license technology. Morton I. Kamien and Israel Zang (1990) also recognize this in studying horizontal mergers.

[3]See Paul R. Milgrom (1985) for an excellent survey.

[4]See Jean Tirole (1989 pp. 350–2) for a detailed discussion.

the cost of preemption as it does not fully incorporate the effect of acquiring an additional unit of capacity on the resulting value of future capacity. His work is discussed in more detail below, in Section III.

David A. Malueg and Marius Schwartz (1991) look at the persistence-of-monopoly issue in a different light. They argue that monopoly may be vulnerable to entry because the incumbent cannot always credibly undertake actions that mimic the outcome with entry. For instance, a monopolist who obtained an additional unit of capacity would not expand output by as much as an entrant would. This phenomenon also occurs in the model presented below.

Also related are papers by Pankaj Ghemawat (1990), Gale (1990b), and Krishna (1990b). Ghemawat examines the value of a single unit of capacity to different sized firms that are engaged in capacity-constrained price competition. He finds that it is more valuable to the larger firm. Although he concludes that this will result in a "snowball effect" whereby the larger firm would keep getting larger, the analysis of this paper suggests that considering the overall game may not lead to such a conclusion. Gale (1990b) looks at the sequential auction of capacity in a Cournot duopoly when the firms may or may not be constrained. When capacity is broken into two units, he shows that both will go to one firm. When it is broken into three units, they need not. This example is related to a general result developed in Krishna (1990b) which shows that when the payoffs to having x of n units is convex in x, monopolization must occur. When two units are sold, the payoffs to having x of n units must be either convex or concave. They tend to be convex because concentration raises industry profits. When three or more units are sold, there is no reason for the payoff function to be convex even if industry profits rise with concentration, so that monopolization need not obtain. An example in Krishna (1990b) also demonstrates that concavity of the payoffs does not imply equal allocations.

This paper is organized as follows. Some of the basic ideas and methodology are introduced in Section I by means of a simple example. The example is that of an abstract, two-object, sequential auction with two bidders. Even though one of the bidders has a higher marginal valuation for both objects than the other, in the unique perfect equilibrium of the auction, he obtains only one of the objects.

Section II introduces a model of a monopoly with a fringe of small competitors and many potential entrants. All firms bid to acquire additional units of a specific capital good, which is referred to as "capacity." As mentioned above, the term "capacity" is quite flexible and could refer to licenses to operate, import, and so on. Two detailed examples in this framework illustrate some of the forces at work. In one, monopoly power dissipates, and in the other, it persists. The difference in the two examples is only in the relative sizes of the "lots" and not in the total capacity sold.

Section III presents two results that generalize the examples of the previous section. Theorem 1 shows that when a number of lots of equal or increasing size become available sequentially, all but the last lot go to the fringe. Thus, monopoly power dissipates. Using this result, I also study some conditions under which capacity divestiture is profitable. Theorem 2 shows that if the units decrease in size at a fast enough rate, all the units go to the monopolist so that monopoly persists. Both results require some assumptions on the shape of the market demand function. Linear demand functions meet these conditions.

In the basic model it is assumed that the capacity-acquisition stage occurs before any production from that capacity takes place. Section IV examines the effect of relaxing this assumption. The timing of payoffs and the discount rate are shown to be crucial to the results. Section V contains some concluding remarks.

I. An Illustrative Example

In this section I present a simple example that illustrates some of the basic issues. Consider a situation in which two agents compete to acquire C units of a resource which will be auctioned off one at a time. I

assume that all the agents have complete information about all aspects of the situation.

Consider first the simplest case in which one object is offered for sale to two agents. In the absence of any uncertainty, the object goes to the agent with the higher valuation, who pays a price (essentially) equal to the other agent's valuation. The exact form of the auction is left unspecified since the outcome would be the same whether the auction were open, a sealed-bid second-price auction, or a sealed-bid first-price auction.[5] Next, suppose that C units are sold sequentially. If the *marginal* valuations of the agents are constant and one agent's marginal valuation is greater than the other's, all the units would be acquired by the agent with the higher valuation.

What if the marginal valuations are not constant? Consider a situation in which two units are sequentially offered to two agents, A and B. Suppose that A assigns a value of 10 to having one unit and a value of 20 to having both. B assigns a value of 9 to having one unit and a value of 10 to having both. To solve for the equilibrium outcome of this game it is necessary to work backwards. This is most easily done by means of Figure 1, in which a tree depicting the possible *outcomes* is drawn. Each unit could go to either A or B. At every node, the left branch denotes that the unit goes to A, and the right branch that it goes to B. It is important to note that this is different from the usual extensive-form depiction of a game. A complete extensive-form description of the game is too complicated. The abbreviated outcome tree contains all the relevant information and is more useful.

The end points of the tree give the payoffs associated with each of the four possible outcomes. Arrows denote the allocation at each stage, and prices are given in brackets next to the paths. Consider the situation when the first unit has gone to A. A is

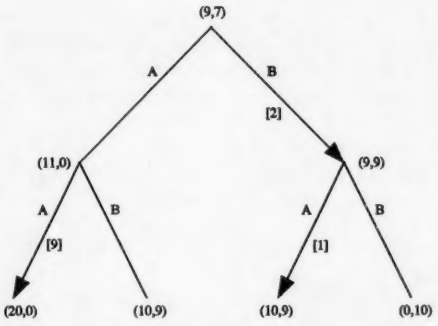


FIGURE 1. DEPICTING EQUILIBRIUM USING AN OUTCOME TREE

willing to pay 10 for the second unit, whereas for B it is worth only 9. Thus, A would obtain the second unit as well, paying a price of 9. If B had the first unit, again the second unit would go to A, but he would pay only 1 for it. How much is A willing to pay for the first unit? To calculate this, the two nodes can be replaced by their equilibrium values. These are (11,0) and (9,9), respectively. The value to A of obtaining the first unit is the difference between the values of the two nodes and thus equals 2. Similarly, the value to B of obtaining the first unit is 9. Thus, B would get the first unit at a price of 2.

Notice that, although A always gets the last unit because his marginal valuation exceeds B's, the first unit goes to B! While at the last stage, the marginal valuation equals the willingness to pay, this is not the case in earlier stages. This is because in earlier stages the willingness to pay is determined by the valuations and the anticipated equilibrium prices in later stages. In the example above, A lets B obtain the first unit so as to get the second unit for a very low price. The example illustrates that, even if one agent's marginal valuations are always greater than those of the other, he need not obtain all the units! This example is at the heart of this paper, as the standard arguments about the persistence of monopoly are usually made in the context of a "once and for all" availability of capacity and rely on the fact that this capacity is more valuable to a monopolist than to an entrant. As

[5]In the first two, the outcome is a result of a dominant-strategy equilibrium. In a first-price sealed-bid auction, it is the result of the a Nash equilibrium in undominated strategies.

I have just shown, when capacity becomes available sequentially this argument is insufficient.

II. A Model of Monopoly with a Fringe

In this section, I outline a simple model of a monopoly with a competitive fringe. It is assumed that there is a large number of potential entrants who can become part of the fringe by obtaining capacity. The firms in the fringe take the price set by the monopolist as the given market price. I wish to study how additional units of capacity, made available sequentially, would be allocated.

Let demand for the product be given by $Q(P)$. Let P^M be the price that would be charged by the monopolist if there were no fringe, and let $Q(P^M) < M$, where M is the capacity that the monopolist already has. Now assume that C additional units of capacity will become available and that having a unit of capacity enables one unit to be produced at a constant marginal cost, assumed without loss of generality to be zero.

If the monopolist obtains x units and the fringe obtains $C - x$ units, the monopolist chooses P to maximize residual profits:

$$(1) \quad \pi(P,x) = P[Q(P) - (C - x)].$$

Let $P(x)$ be the profit-maximizing price and let $f(x) = \pi(P(x), x)$ denote the maximized value of the monopolist's profits. Note that $P(x)$ rises with x assuming that $\pi(P,x)$ is concave in P given x. Assume $P(0) > 0$. Also, note that $f'(x) = P(x)$, by the envelope theorem. These properties imply that the marginal value of an additional unit of capacity to the monopolist is positive and increasing.

The standard argument for persistence of monopoly is easy to see from Figure 2, which depicts $P(x)$. Suppose that all C units became available as a lump. The value of this capacity to the monopolist is $f(C) - f(0)$. By the fundamental theorem of calculus, this equals

$$\int_0^C P(t)\,dt$$

which is the area under the $P(\cdot)$ curve.

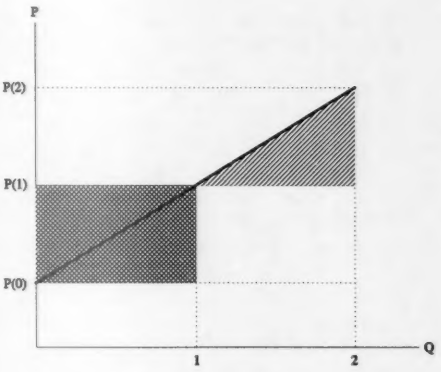


FIGURE 2. AUCTION OF CAPACITY IN TWO LOTS OF EQUAL SIZE

The value of this capacity to an entrant, assuming he had none to begin with, is $P(0)C$, since if he acquired the capacity, the price would be $P(0)$.[6] From Figure 2 it is clear that the value of the additional capacity and hence the willingness to pay for it is greater for the monopolist than for an entrant. Thus, the additional capacity will go to the monopolist. This is the essence of the argument for the persistence of monopoly given by Gilbert and Newbery (1982), among others.

To begin, consider a situation in which the two units of additional capacity are sequentially auctioned off. Suppose the demand curve is linear, so that $P(\cdot)$ is also linear, as drawn in Figure 2. Assume that there are $R > 2$ potential entrants so that the situation is a game with $R + 1$ players.

[6] If the entrant had any capacity, his willingness to pay would be less than that of a new entrant. A new entrant's valuation is the value of the units in the equilibrium of the subgame in which he gets the units in question. The willingness to pay of an entrant with some capacity is the difference between his total payoff when he gets the unit(s) and when he does not. This must be weakly less than that of a new entrant, as the price charged falls as the monopolist gets fewer units, which reduces the value of units already held. Thus, existing fringe members are outbid by new entrants. This also ensures that no fringe-member firm grows large enough not to use all its capacity.

Suppose that the first unit goes to the monopolist. Then the monopolist is willing to pay $f(2)-f(1)$ for the second unit. On the other hand, the entrant is willing to pay $P(1)1$ as $P(1)$ is the price that would occur if the entrant got the last unit. Again, note that the monopolist's willingness to pay for the second unit, which equals the area under the $P(\cdot)$ line, exceeds that of the entrant, since $P(\cdot)$ is an increasing function. Hence, in this subgame, the last unit goes to the monopolist who pays $P(1)$. If the first unit goes to an entrant, the monopolist is willing to pay $f(1)-f(0)$ for the second unit, whereas a new entrant is willing to pay $P(0)$. As the willingness to pay of a new entrant exceeds that of an existing fringe firm, only new entrants need to be considered. This occurs because a fringe firm with some capacity would internalize the fall in price from the monopolist not getting the unit, while a new entrant would not (see footnote 6). Thus, the monopolist's willingness to pay exceeds that of a new entrant, so that the last unit goes to the monopolist who pays $P(0)$ for it.

Now consider the first unit. If the monopolist gets it, his payoff is $f(2)-P(1)1$. If he does not, his payoff is $f(1)-P(0)1$. Thus his willingness to pay is $[f(2)-f(1)]-[P(1)-P(0)]1$. If an entrant gets the first unit, he knows that the second will go to the monopolist, so that the resulting price will be $P(1)$. Thus, an entrant will be willing to pay $P(1)1$ for the first unit. With linear demand, the monopolist's willingness to pay is less than that of an entrant since $[f(2)-f(1)-P(1)]-[P(1)-P(0)]1<0$. The first term equals the striped triangle of Figure 2, while the second is the cross-hatched rectangle, which has an area twice that of the triangle. In equilibrium, therefore, an entrant gets the first unit while the monopolist gets the second, and monopoly is eroded! The reason is the same as in the example of the previous section: obtaining the first unit makes it costly to obtain the second, and the monopolist is better off letting it go to an entrant.

To get further intuition about this, consider the more general situation in which C units of capacity are sequentially auctioned off in two lots. The first lot is of size x and the second is of size $C-x$.[7] Once again, it is easy to verify that the last lot (of size $C-x$) always goes to the monopolist. If the first lot (of size x) went to the entrant, the price paid for the last lot would be $P(0)(C-x)$, as $P(0)$ is the price that would prevail if the entrant got the last lot as well. If the first lot went to the monopolist, the price paid for the second lot would be $P(x)(C-x)$ for similar reasons. Thus, the payoff to the monopolist if the first lot went to the entrant is $f(C-x)-P(0)(C-x)$, and the price in this subgame is $P(C-x)$. The entrant's willingness to pay for the first lot is therefore $P(C-x)x$. If the first lot went to the monopolist, his payoff would be $f(C)-P(x)(C-x)$, as the last lot always goes to the monopolist.

Now look at the first stage. The difference between the willingness to pay of the monopolist and that of the entrant for the first lot is

$$(2)\quad [f(C)-f(C-x)-P(C-x)x]-[P(x)-P(0)](C-x).$$

The first term in brackets is positive and reflects the higher willingness to pay of the monopolist at the last stage. However, if he gets the first lot, he pays more for the $C-x$ units in the second stage. This is captured by the second term. With linear demand, the above expression has the same sign as $x-2(C-x)$. Hence, if $x>2(C-x)$, the monopolist would get *both* lots. If, on the other hand, $x<2(C-x)$, the entrant would get the first lot. The intuition is that if x is large, the gain from having a lower price for the second lot is small, and the game resembles the single-stage game. In the next section, the relationship between the sizes of the lots and the persistence of monopoly is explored in more detail.

[7] The reader may find it useful to draw the outcome tree corresponding to this problem.

III. Two General Results

Suppose that C units of additional capacity are made available in n lots which are sold sequentially. The size of the kth lot is x_k. Theorem 1 shows that if $P(\cdot)$ is concave and if for all k, $x_k \leq x_{k+1}$, then the monopolist wins the auction only in the last stage, so that monopoly power is dissipated. This is quite a strong result: despite having a higher marginal valuation than entrants, the monopolist chooses to let entrants obtain all but the last lot. Theorem 2 shows that if demand is linear and for all k, $x_k > 2X_{k+1}$, where $X_k = \sum_{j=k}^{n} x_j$, then monopoly persists. Note that this requires x_k to fall very fast.

The analysis is facilitated by defining a sequence of functions that give the monopolist's equilibrium payoffs as obtained by folding the outcome tree back. For $k = 0, 1, 2, \ldots, n$, define $f_{n-k}(x)$ to be the monopolist's equilibrium payoff in the subgame in which he has x units of capacity and the last k lots remain. By definition, $f_n(x) = f(x) \equiv \pi(P(x), x)$. For $k > 0$, $f_{n-k}(x)$ is the larger of the values from the two options open to the monopolist. The monopolist could get the $(n-k+1)$st lot at the price he would have to pay for it, which is the same as the market price that would prevail in the subgame in which an entrant obtained the $(n-k+1)$st lot. The other option is to let the entrant get the lot. This defines the function recursively. While this is difficult to evaluate in general, Lemmas 1 and 2 evaluate it for two interesting cases.

LEMMA 1: *If $P(\cdot)$ is concave and if for all $k, x_k \leq x_{k+1}$, then*

$$(3)\quad f_{n-k}(x) = f_{n-k+1}(x) \qquad \text{for } k > 1$$

$$= f_{n-k+1}(x + x_n) - P(x)x_n \qquad \text{for } k = 1.$$

so that by repeated application, for all k,

$$(4)\quad f_{n-k}(x) = f_n(x + x_n) - P(x)x_n.$$

(The proof is contained in the Appendix.)

Theorem 1 below is an immediate consequence of Lemma 1 and provides sufficient conditions for the dissipation of monopoly power.

THEOREM 1: *If $P(\cdot)$ is concave and for all k, $x_k \leq x_{k+1}$, then in equilibrium:*

(i) the first $n-1$ lots go to new entrants while only the last lot goes to the monopolist;
(ii) the price per unit of capacity paid by entrants equals $P(x_n)$, while the monopolist pays $P(0)$ per unit of capacity; and
(iii) entrants earn zero profits ex ante while the monopolist earns profits of $f_n(x_n) - f_n(0) - P(0)x_n > 0$ from participating in the auction.

PROOF:

Part (i) follows directly from Lemma 1. For part (ii), by Lemma 1 the equilibrium allocation in every subgame consists of all but the last unit going to an entrant. The monopolist pays an entrant's willingness to pay, and the entrant pays the equilibrium price in the subgame where he obtains the unit, which are the prices given. Part (iii) follows as in every subgame entrants pay the full value of the object in equilibrium, and the monopolist only pays the entrant's marginal valuation at the last stage.

Lemma 1 and Theorem 1 can also be used to shed some light on a few related questions.

First, consider the problem of finding the revenue-maximizing lot size in a sequential auction. Although the general problem is rather difficult, restricting attention to equal-lot-size auctions makes it tractable. Suppose that the total capacity, C, is sold in n lots of size x each and also suppose that demand is linear. By Theorem 1, the revenue raised in a sequential auction is calculated as follows:

$$(5)\quad R(x) = P(x)(n-1)x + P(0)x$$

$$= P(x)xn - [P(x) - P(0)]x$$

$$= P(x)C - [P(x) - P(0)]x.$$

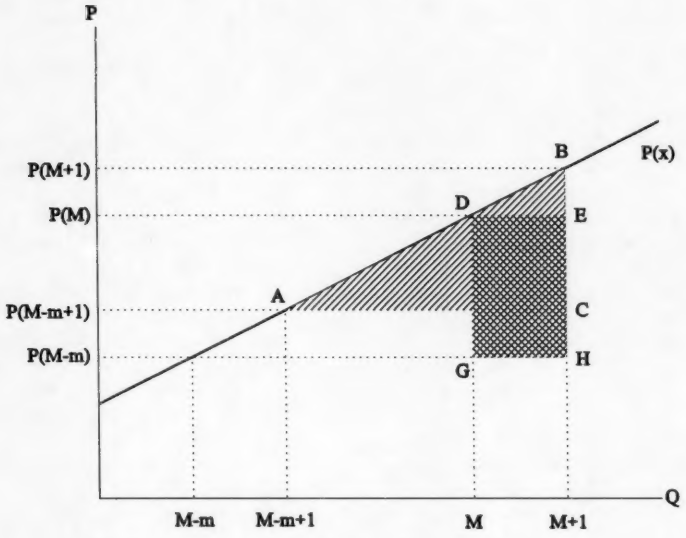


FIGURE 3. DIVESTITURE OF CAPACITY

Since $P(\cdot)$ is linear, $R(x)$ is maximized at $x = C/2$ where C equals the total capacity that is to be auctioned off.

The case of equal-sized lots was also examined by Lewis (1983). Without loss of generality, assume that each lot consists of one unit of capacity. Lewis looks at the cost of acquiring k of n units and argues that the price the monopolist must pay to do so is $P(k)k$, as this is the value of k units if he succeeds in doing so. Reexamining this in the context of a sequential auction as above, one sees that if the monopolist wanted to get k of n units available, he would have to pay $P(j)$ for the jth unit he obtained ($j < k$). This is because in the subgame in which the entrant gets the unit in question, only the last unit would go to the monopolist, as in Lemma 1. Thus, if the monopolist had $j-1$ units and the entrant obtained the next unit, the price in the subgame would be $P(j)$, which is the entrant's valuation and the price the monopolist would have to pay. The monopolist always gets the last unit, and if he already had $k-1$ units he would have to pay only $P(k-1)$ for this. The cost of acquiring k units is thus less than $P(k)k$.

Lewis also argues, on the basis of looking at the cost of preemption alone, that divestiture of capacity may be an attractive proposition for a firm. His basic insight is that divestiture reduces the price the firm has to pay for preemptively acquiring any given amount of capacity. However, this is again only part of the story. Assume, as before, that the firm has an initial capacity of M. If it chooses to divest itself of m units of capacity it enters a subgame in which it has $M-m$ units and $n+m$ units of capacity are available. Figure 3 serves as a guide to the discussion below.

In the equilibrium of the game, assuming it divests, the monopolist gets only one unit of capacity and pays $P(M-m)$ for it.[8] Entrants pay $P(M-m+1)$ for each unit. Thus the monopolist obtains $mP(M-m+1)$ in revenue from the sale of m units of capacity. Finally, as the monopolist firm's capacity in equilibrium is $M-m+1$, it gets profits of $f(M-m+1)$. If it does not divest, it gets profits of $f(M+1)$ and pays $P(M)$ for

[8]Note that $P(M-m)$ corresponds to the monopolist getting none of the $m+n$ units available.

the capacity it does obtain. The difference in the payoff from not divesting and divesting equals

$$(6)\qquad [f(M+1)-f(M-m+1) - P(M-m+1)m] - [P(M)-P(M-m)].$$

Divesting reduces the profits *ex post*, as the monopolist firm has less of the total capacity and so charges a lower price and gets lower profits. However, it also obtains the revenues from the sale of capacity. The net effect of these is given by the first term in brackets. This is positive as $P(\cdot)$ is an increasing function. This is evident from Figure 3, where it equals the striped triangle, ABC.

However, the monopolist firm pays a lower price for the one unit it purchases when it divests. This is the second term in the above expression. This term equals the area of the cross-hatched rectangle DEGH in Figure 3. While the first term works against divestiture, the second term works for it. The total effect could be positive or negative.

If $m=1$ and $P(\cdot)$ is concave, then the gain from not divesting equals

$$(7)\quad [f(M+1)-f(M)-P(M)] - [P(M)-P(M-1)] = \{f(M+1)-f(M)-P(M+1)\} + \{[P(M+1)-P(M)] - [P(M)-P(M-1)]\}.$$

If $P(\cdot)$ is concave, both terms are negative, and divestiture is profitable.

If $P(\cdot)$ is linear and m units are divested, the first term in the general expression above equals $\frac{1}{2}\alpha m^2$, while the second equals $m\alpha$ where α is the slope of $P(\cdot)$. Their difference equals $m\alpha[(m/2)-1]$. For $m>2$, this is positive so that divestiture reduces profits if more than two units are divested; thus the potential for raising profits by divestiture is relatively limited.

While Theorem 1 and its applications all concern the dissipation of monopoly power, I now turn to an examination of some conditions under which monopoly would persist. For the two-stage example in the previous section, I showed that if $x>2(C-x)$ the monopolist obtained both lots. For *all* units to go to the monopolist in the n-stage game, a generalization of this condition is needed. This is the content of Lemma 2 and Theorem 2.

The relationship between $f_{n-k}(x)$ and $f_n(x)$ that would result if the monopolist were to obtain all additional units in every subgame is derived in Lemma 2. Theorem 2 uses this relationship to determine the equilibrium outcomes.

LEMMA 2: *If P is linear and, for all k,*

$$x_k > 2X_{k+1} \qquad \text{where } X_k = \sum_{i=k}^{n} x_i$$

then for all k,

$$(8)\quad f_{n-k}(x) = f_{n-k+1}(x + x_{n-k+1}) - P\left(x + \sum_{i=2}^{k} x_{n-k+i}\right)x_{n-k+1}.$$

(See the Appendix for the proof.)

THEOREM 2: *If $P(\cdot)$ is linear and for all k, $x_k > 2X_{k+1}$, then in equilibrium:*

(i) all units go to the monopolist;
(ii) the price per unit paid in stage k equals $P(X_1 - x_k)$; and
(iii) the monopolist's profits from participating in the auction, $f_0(0) - f_n(0)$, are positive.

PROOF:

Parts (i) and (ii) follow from Lemma 2. Part (iii) then follows from Lemma 2 and the fact that at any stage, and following any history, it is profitable for the monopolist to

obtain the current lot. Hence,

$$(9)\quad f_0(0) > f_1(0) > f_2(0) > \ldots > f_n(0).$$

While the model above has been formulated in terms of capacity acquisition, it affords another interpretation. Consider the problem of a monopolist who operates in more than one country. It is convenient to think of the monopolist as being foreign and operating in the home market and in another market. Arbitrage forces him to charge one price for his product. Let $Q(P)$ denote the demand at home. If he charges P he gets $Q(P)P$ revenue from the home market and $r(P)$ from his other markets. Suppose that the home country imposes a quota at the level of imports denoted by V. If the monopolist charges P, then the price of a license to import one unit is $L(P,V)$, where $L(\cdot)$ is defined by $Q(P+L)=V$. Now suppose that these licenses are auctioned to the monopolist or to a large number of competitive domestic retailers.[9] If the monopolist gets x of these licenses, he chooses a price to maximize:

$$(10)\quad \pi(P,x,V) = PQ(P+L(\cdot)) + r(P) + xL(\cdot).$$

Denote this price by $P(x,V)$ and let $\pi(P(x,V),x,V)=f(x,V)$. Note that by the envelope theorem $f'(x,V)=L(P(x,V),V)$, which also equals a competitive retailer's valuation of a license.[10] For fixed V, this model has the same structure as the model of monopoly with a fringe.

IV. The Importance of Timing

The model studied implicitly assumes that the acquisition stage occurs before any production takes place. Thus, the acquisition stage is assumed to be brief compared to the life of the industry. This may be more appropriate in some settings (e.g., the quota license interpretation offered above) than others. In this section, I examine the effects of relaxing this assumption.

The main issues can be seen by amending the simple two-stage example of Section II. As before, in the presence of a competitive fringe the monopolist is assumed to have enough capacity to produce whatever he chooses. As before, in each period one unit of capacity becomes available and can be acquired by either the monopolist or an entrant. However, production is now assumed to take place in *each* of the two periods. Thus, the value of winning the first unit of capacity now includes the profits from production in the first period.

Let A_{11} be the second-period profits accruing to the monopolist when he has both units of additional capacity. Similarly, let A_{10} be his second-period profits from having the first but not the second, let A_{01} be the profits from having the second but not the first, and let A_{00} be the result of having neither. Having obtained the first unit, the monopolist gets $f(2)$ in the second period if he obtains the second unit. Hence, $A_{11}=f(2)$. If he does not obtain the second unit, he only gets $f(1)$, so $A_{10}=f(1)$. Similarly, $A_{01}=f(1)$, and $A_{00}=f(0)$.

The entrant's bid for the second unit of capacity when the monopolist gets the first unit is $P(1)$, and the bid is $P(0)$ when the monopolist does not. It is easy to see that, as in the previous sections, the monopolist gets the last unit in both subgames.

However, incentives in the *first* stage are now different. Let A_1 denote the monopolist's payoff evaluated at the beginning of the first stage when he gets the first unit. Let A_0 be similarly defined as his profits when he does not get the first unit. If the monopolist gets the first unit, he gets $f(2)$ from production in the first period (as no capacity has gone to the fringe) and $\delta[f(2)-P(1)]$ in present-value terms from the subgame, where $\delta=1/(1+r)$ is the discount factor. If he does not get the first unit he gets $f(1)$ from production in the first period (as one unit of capacity goes to the

[9] If there are only a few retailers, then the model would not apply, as the marginal valuation of the retailers would lie *below* that of an entrant and be less than $L(\cdot)$ because retailers would take into account how getting more licenses would reduce the value of the licenses they already own.

[10] Generally $L(\cdot)$ is positive, as assumed here, if the quota is relatively restrictive.

fringe) and $\delta[f(1)-P(0)]$ in present-value terms from the second period. An entrant is willing to pay $P(1)(1+\delta)$ for this capacity, as the monopolist gets the second unit, and the price is $P(1)$ throughout. The difference between the willingness to pay of the monopolist and the entrant's bid equals

$$(11)\quad f(2)-f(1)-P(1)+\delta\{f(2)-f(1)-P(1)-[P(1)-P(0)]\}.$$

In the linear case, this equals

$$(12)\qquad (1-\delta)\alpha/2>0$$

where, as before, α is the slope of the $P(\cdot)$ function.

Hence, the monopolist gets *both* units! It is easy to verify that, if production (but not capacity acquisition) continues after the second period, the last unit would go to the monopolist. However, if $r<1$, the first unit would go to the entrant.[11]

When production occurs before the acquisition stage ends, acquiring capacity today raises profits today as well as raising the price of capacity acquisition in the future. The former works in favor of monopoly persisting, while the latter works against it. If r is small, the latter predominates, and monopoly power dissipates; while if r is large, the former predominates, and monopoly persists.

V. Conclusion

I have argued that the sequential nature of the competition to acquire resources can have a powerful effect on their resulting allocation. In particular, even though large incumbent firms attach greater value to these resources than new entrants, some of these additional resources may go to entrants. The basic insight is that the prices of later units rise if earlier units are acquired.

If the capacity-acquisition stage is short and unit size rises weakly, then monopoly dissipates. This also occurs if the capacity-acquisition stage is long but the future matters a lot, either because the interest rate is low or because the life of the industry is long relative to the acquisition stage. On the other hand, if the capacity-acquisition stage is short, preemption occurs if the unit size falls sharply enough. If the capacity-acquisition stage is long relative to the life of the industry, preemption occurs if the future matters little, either because of discounting or because production ends with period 2. Thus, the issue of the persistence of monopoly depends critically on the sequence of lot size, the timing of capacity acquisition and production, and on the discount factor.

[11] The last unit goes to the monopolist as before. The first unit also goes to the monopolist if the monopolist's willingness to pay at the first stage exceeds that of the entrant. The difference in the two has the same sign as $[1-(1/r)]\alpha/2$.

Appendix

LEMMA 1: *If $P(\cdot)$ is linear or concave and if for all k, $x_k \le x_{k+1}$, then*

$$(A1)\quad f_{n-k}(x)=f_{n-k+1}(x)\qquad \text{for } k>1$$
$$=f_{n-k+1}(x+x_n)-P(x)x_n\qquad \text{for } k=1$$

so that by repeated application, for all k,

$$(A2)\quad f_{n-k}(x)=f_n(x+x_n)-P(x)x_n.$$

PROOF:

For $k=1$,

$$(A3)\quad f_{n-1}(x)=\max\begin{cases}f_n(x)\\ f_n(x+x_n)-P(x)x_n\end{cases}$$
$$=f_n(x+x_n)-P(x)x_n.$$

For $k>1$, the proof is by induction. For $k=2$,

$$(A4)\quad f_{n-2}(x)=\max\begin{cases}f_{n-1}(x)\\ f_{n-1}(x+x_{n-1})\\ \quad -P(x+A(x,n-1))x_{n-1}\end{cases}$$

where $A(x, n-k)$ is the allocation, that is, the amount of new capacity obtained by the monopolist, in the subgame at the $(n-k)$th stage when the monopolist already has x units. Moreover,

$$\text{(A5)}\quad f_{n-2}(x) = \max\begin{cases} f_n(x+x_n) - P(x)x_n \\ f_n(x+x_{n-1}+x_n) \\ \quad - P(x+x_{n-1})x_n \\ \quad - P(x+x_n)x_{n-1} \end{cases}$$

since $A(x, n-1) = x_n$, as the monopolist always wins in the last stage, and as Lemma 1 holds for $k=1$. The lower term, L, is less than the upper one, U, as their difference $L-U$ is

$$\begin{aligned}\text{(A6)}\quad & \{f_n(x+x_{n-1}+x_n) - f_n(x+x_n) \\ & \qquad - P(x+x_{n-1}+x_n)x_{n-1}\} \\ & + \left[\frac{P(x+x_{n-1}+x_n) - P(x+x_n)}{x_{n-1}}\right] x_{n-1}^2 \\ & - \left[\frac{P(x+x_{n-1}) - P(x)}{x_{n-1}}\right] x_n x_{n-1} \\ & \le \{f_n(x+x_{n-1}+x_n) - f_n(x+x_n) \\ & \qquad - P(x+x_{n-1}+x_n)x_{n-1}\} \\ & + \left[\frac{P(x+x_{n-1}+x_n) - P(x+x_n)}{x_{n-1}}\right] x_{n-1}x_n \\ & - \left[\frac{P(x+x_{n-1}) - P(x)}{x_{n-1}}\right] x_{n-1}x_n \\ & < 0.\end{aligned}$$

The first inequality occurs because $x_{n-1} \le x_n$. The second occurs because the first term is negative as $P(\cdot)$ is upward-sloping and the difference between the second and third terms is nonpositive if $P(\cdot)$ is concave.

Finally, assuming that the result holds for k, I show that it holds for $k+1$. Again,

$$\text{(A7)}\quad f_{n-(k+1)}(x) = \max\begin{cases} f_{n-k}(x) \\ f_{n-k}(x+x_{n-k}) \\ \quad - P(x+A(x, n-k))x_{n-k}. \end{cases}$$

As the result holds for k by assumption, $A(n-k, x) = x_n$, and using the result for k repeatedly gives

$$\text{(A8)}\quad f_{n-(k+1)}(x) = \max\begin{cases} f_n(x+x_n) - P(x)x_n \\ f_n(x+x_{n-k}+x_n) \\ \quad - P(x+x_{n-k})x_n \\ \quad - P(x+x_n)x_{n-k}. \end{cases}$$

Again, $L-U$ equals

$$\begin{aligned}\text{(A9)}\quad & \{f_n(x+x_{n-k}+x_n) - f(x+x_n) \\ & \qquad - P(x+x_{n-k}+x_n)x_{n-k}\} \\ & + \left[\frac{P(x+x_{n-k}+x_n) - P(x+x_n)}{x_{n-k}}\right] x_{n-k}^2 \\ & - \left[\frac{P(x+x_{n-k}) - P(x)}{x_{n-k}}\right] x_n x_{n-k} \\ & \le \{f_n(x+x_{n-k}+x_n) - f(x+x_n) \\ & \qquad - P(x+x_{n-k}+x_n)x_{n-k}\} \\ & + \left[\frac{P(x+x_{n-k}+x_n) - P(x+x_n)}{x_{n-k}}\right] x_{n-k}x_n \\ & - \left[\frac{P(x+x_{n-k}) - P(x)}{x_{n-k}}\right] x_{n-k}x_n \\ & < 0.\end{aligned}$$

The first inequality holds as $x_{n-k} \le x_n$. The second holds as the first term is negative and the difference between the last two is nonpositive since $P(\cdot)$ is concave. Thus, the induction step holds.

LEMMA 2: *If P is linear and, for all k,*

$$x_k > 2X_{k+1}$$

where $X_k = \sum_{i=k}^{n} x_i$ then for all k,

$$\text{(A10)}\quad f_{n-k}(x) = f_{n-k+1}(x + x_{n-k+1}) - P\left(x + \sum_{i=2}^{k} x_{n-k+i}\right)x_{n-k+1}.$$

PROOF:

The last unit always goes to the monopolist so that this holds for $k = 1$ since for $k > n$, $x_k = 0$. Thus, $A(x, n-1) = x_n$. When $k = 2$,

$$\text{(A11)}\quad f_{n-2}(x) = \max\begin{cases} f_{n-1}(x) \\ f_{n-1}(x + x_{n-1}) - P(x + A(x, n-1))x_{n-1} \end{cases}$$

$$= \max\begin{cases} f_n(x + x_n) - P(x)x_n \\ f_n(x + x_{n-1} + x_n) - P(x + x_{n-1})x_n - P(x + x_n)x_{n-1}. \end{cases}$$

The difference between the lower and upper terms, $L - U$, is given by

$$\text{(A12)}\quad L - U = \{f_n(x + x_{n-1} + x_n) - f_n(x + x_n) - P(x + x_n)x_{n-1}\} - [P(x + x_{n-1}) - P(x)]x_n.$$

As $P(\cdot)$ is linear with slope α,

$$\text{(A13)}\quad L - U = \frac{1}{2}(x_{n-1})^2\alpha - \alpha x_{n-1}x_n = \frac{\alpha x_{n-1}}{2}[x_{n-1} - 2x_n] > 0$$

as by assumption $x_{n-1} > 2x_n$. Thus, $f_{n-2}(x) = f_{n-1}(x + x_{n-1}) - P(x + x_n)x_{n-1}$ as postulated.

Now I show that, if the result holds for k, it holds for $k+1$. Note first that

$$\text{(A14)}\quad \sum_{i=2}^{k} x_{n-k+i} = \sum_{i=n-k+2}^{n} x_i = X_{n-k+2}.$$

As the result holds for k by assumption,

$$\text{(A15)}\quad f_{n-k}(x) = f_n(x + X_{n-k+1}) - \sum_{i=1}^{k} [P(x + X_{n-k+1} - x_{n-k+i}) \times x_{n-k+i}].$$

Similarly,

$$\text{(A16)}\quad f_{n-k}(x + x_{n-k}) = f_n(x + x_{n-k} + X_{n-k+1}) - \sum_{i=1}^{k} [P(x + x_{n-k} + X_{n-k+1} - x_{n-k+i})x_{n-k+i}].$$

Thus,

$$\text{(A17)}\quad f_{n-(k+1)}(x) = \max\begin{cases} f_{n-k}(x) \\ f_{n-k}(x + x_{n-k}) - P(x + X_{n-k} - x_{n-k})x_{n-k} \end{cases}$$

$$= \max\begin{cases} f_n(x + X_{n-k+1}) - \sum_{i=1}^{k} P(x + X_{n-k+1} - x_{n-k+i})x_{n-k+i} \\ f_n(x + X_{n-k}) - \sum_{i=1}^{k} P(x + X_{n-k} - x_{n-k+i})x_{n-k+i} - P(x + X_{n-k} - x_{n-k})x_{n-k}. \end{cases}$$

The lower term less the upper one is de-

noted by $L-U$, and

$$\text{(A18)}\quad L-U = \{f_n(x+X_{n-k}) - f_n(x+X_{n-k+1}) - P(x+X_{n-k+1})x_{n-k}\} + \sum_{i=1}^{k} [P(x+X_{n-k}-x_{n-k+i}-x_{n-k}) - P(x+X_{n-k}-x_{n-k+i})]x_{n-k+i}.$$

As $P(\cdot)$ is linear,

$$\text{(A19)}\quad L-U = \frac{1}{2}\alpha(x_{n-k})^2 - \alpha x_{n-k}\sum_{i=1}^{k} x_{n-k+i} = \frac{\alpha x_{n-k}}{2}[x_{n-k}-2X_{n-k+1}] > 0$$

as, for all k, $x_{n-k} > 2X_{n-k+1}$.

Thus, the result holds for $k+1$ if it holds for k.

REFERENCES

Bergsten, C. F., Elliot, Kimberly A., Schott, Jeffrey J. and Takacs, Wendy E., "Auction Quotas and United States Trade Policy," in *Policy Analyses in International Economics*, Vol. 19, Washington, DC: Institute for International Economics, 1987, pp. 173–82.

Feenstra, Robert C., Lewis, Tracy R. and MacMillan, John, "Designing Policies to Open Trade," National Bureau of Economic Research (Cambridge, MA) Working Paper No. 3258, 1990.

Gale, Ian, (1990a) "A Multiple Object Auction with Superadditive Values," *Economics Letters*, 1990, *34* (4), 323–8.

_____, (1990b) "On the Use of Competitive Markets to Allocate Industry Capacity," mimeo, University of Wisconsin, 1990.

Ghemawat, Pankaj, "The Snowball Effect," *International Journal of Industrial Organization*, September 1990, *8*, 335–52.

Gilbert, Richard J. and Newbery, David M. G., "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

Kamien, Morton I. and Zang, Israel, "The Limits of Monopolization through Acquisition," *Quarterly Journal of Economics*, May 1990, *105*, 465–99.

Katz, Michael L. and Shapiro, Carl, "How To License Intangible Property," *Quarterly Journal of Economics*, August 1986, *101*, 567–89.

Krishna, Kala, (1990a) "The Case of the Vanishing Revenue: Auction Quotas with Monopoly," *American Economic Review*, September 1990, *80*, 828–36.

_____, (1990b) "Auctions With Endogenous Valuations, the Snowball Effect, and Other Applications," National Bureau of Economic Research (Cambridge, MA) Working Paper No. 3483, 1990.

Lewis, Tracy R., "Preemption, Divestiture, and Forward Contracting in a Market Dominated by a Single Firm," *American Economic Review*, December 1983, *73*, 1092–1101.

Malueg, David A. and Schwartz, Marius, "Preemptive Investment, Toehold Entry, and the Mimicking Principle," *Rand Journal of Economics*, Spring 1991, *22*, 1–13.

Milgrom, Paul R., "The Economics of Competitive Bidding: A Selective Survey," in Lionel Hurwicz, David Schmeidler, and Hugo Sonnenschein, eds., *Social Goals and Social Organizations*, Cambridge: Cambridge University Press, 1985, pp. 261–89.

Pitchik, Carolyn, "Budget Constrained Sequential Auctions with Incomplete Information," Suntory Toyota International Center for Economics and Related Disciplines Working Paper TE/89/201, London School of Economics, 1989.

Tirole, Jean, *The Theory of Industrial Organization*, Cambridge, MA: MIT Press, 1989, pp. 350–2.

A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand

By Richard Arnott, André de Palma, and Robin Lindsey*

This paper considers the modeling of road congestion subject to peak-load demand. The standard model contains ambiguities and is poorly specified. These problems can be eliminated by working with a structural model that explicitly treats the congestion technology and drivers' behavioral decisions. The paper provides a detailed analysis of a particular structural model—William Vickrey's model of bottleneck congestion in the morning rush-hour auto commute, extended to treat elastic (i.e., price-sensitive) demand—and examines some economic implications of the structural approach. (*JEL* R41)

This paper considers the modeling of road congestion subject to peak-load demand. It argues that a properly specified model should be *structural*, that is, be derived explicitly treating the congestion technology and consumers' behavioral decisions.

The standard model of a facility subject to peak-period congestion is specified as follows (e.g., Dennis J. Aigner and Joseph G. Hirschberg, 1985; Ronald R. Braeutigam, 1989).[1] The period of use is divided into finite time intervals. A separate analysis is conducted in price–quantity space for each of these intervals. The quantity is the number of facility users. Price is the sum of the user's toll or fee and his private cost (net of the toll, but incorporating congestion). Marginal social cost is the increase in social cost in that interval from an additional user in that interval. Both private and marginal social cost are viewed as depending on the number of users and capacity and as being determined by the technology. The position of the demand curve depends on both the time interval and the equilibrium prices for the other time intervals (because of intertemporal substitution). Equilibrium is at the point of intersection of the demand and supply (relating price to the number of facility users) curves, and the social optimum at the point of intersection of the demand (marginal-social-benefit) and marginal-social-cost curves.

This specification contains a number of ambiguities. To illustrate, consider a stylized model of traffic in the morning rush hour in which congestion occurs at only a single bottleneck, the entry point to the central business district. This bottleneck has fixed capacity, and if the number of drivers arriving at the bottleneck exceeds this capacity a queue forms. Even for this simple

*Arnott: Department of Economics, Boston College, Chestnut Hill, MA, 02167-3806; de Palma: Department of Civil Engineering, Northwestern University, Evanston, IL, 60208-3109; Lindsey: University of Alberta, Edmonton, Alberta, Canada, T6G 2H4. Financial support from NSERC, NATO, and NSF SES-8912335 is gratefully acknowledged. We also thank seminar participants at Universitat Autonoma de Barcelona, Calgary, Harvard, Laval, Princeton, Saskatchewan, and Trent universities, the London School of Economics, and the Séminaire René Roy, Paris, as well as Marvin Kraus, for helpful comments. This paper was redrafted while the authors were visiting GREQE, Marseilles; we thank the group there for its hospitality.

[1] Early works on peak-load pricing include Jules Dupuit (1849), Arthur C. Pigou (1920), and Frank Knight (1924); all of these authors recognized the application of the principle of marginal-cost pricing to traffic congestion. Major contributions to the modern theory of peak-load pricing include Marcel Boiteux (1949), Peter O. Steiner (1957), Oliver E. Williamson (1966), and Michael A. Crew and Paul R. Kleindorfer (1986). Important papers applying the theory to urban transportation include Alan A. Walters (1961), Herbert D. Mohring and Mitchell Harwitz (1962), Robert H. Strotz (1965), Marvin C. Kraus et al. (1976), and Theodore E. Keeler and Kenneth A. Small (1977).

example, there are several respects in which it is not clear how to apply the standard model. First, what is the appropriate measure of the number of users in a time interval? Is it the number of drivers joining the queue or the number passing through the bottleneck? This ambiguity stems from a failure to model the congestion technology explicitly. Second, the private cost in an interval depends not only on the number of users in that interval, but also on queue length, which reflects congestion in previous intervals. This illustrates a point that transport engineers have stressed: that congestion is inherently a dynamic phenomenon. Third, and relatedly, the addition of a driver in an interval increases the queue length and hence social cost not just in that interval, but in other intervals as well. This list of ambiguities suggests that the standard model of peak-load congestion in which the period of use is divided into intervals is poorly specified.

In recent years a new structural model of peak-period congestion that does not contain these ambiguities has been extensively developed. The main innovation is to treat explicitly the user's behavioral decisions. The literature has focused on the time-of-use decision, whereby the user trades off the cost of using the facility at an inconvenient time against the congestion cost of using the facility when it is crowded. The structural approach also provides an explicit treatment of the congestion technology.

The seminal paper is one by William Vickrey (1969), and the example we gave above is based on his model. He examined equilibrium with a fixed number of identical drivers on a point-input, point-output road in the morning rush hour, along which there is a single bottleneck of fixed capacity. If the arrival rate at the bottleneck exceeds capacity, a queue develops. All drivers wish to arrive at work at the same time. (Actually, Vickrey [1969] considered a situation with a distribution of work start times, but the bulk of the literature assumes a common start time.) This is physically impossible. As a result, each driver faces a trade-off in deciding when to leave home. If she leaves early, she faces no queue but arrives at work inconveniently early; if she leaves so as to arrive on time, she faces a long queue; and if she leaves late, she faces no queue, but arrives inconveniently late. Equilibrium obtains when the queue length over time is such that no driver can reduce her trip price by changing her departure time. Vickrey not only solved for the no-toll equilibrium, but also determined the social optimum and the time-varying toll which decentralizes it and made some astute observations concerning optimal capacity and equilibrium with two parallel roads.

The Vickrey (1969) model has been elaborated in a large number of papers in transportation science (e.g., Chris Hendrickson and George Kocur, 1981), but with the exception of papers by the present authors (e.g., de Palma and Arnott, 1986; Arnott et al., 1989, 1990; Arnott and Kraus, 1990) and by Ralph M. Braid (1989) and Yuval Cohen (1987), it has been ignored by economists.[2] This is unfortunate since the rich economic implications of the model have not received the attention they merit. In this paper, we extend the model, focusing on its economics. Our paper makes four contributions to the literature on peak-period congestion.

First, the paper illustrates the power of the structural approach in analyzing alternative pricing regimes and provides the first treatment of elastic (i.e., price-sensitive) trip demand in the context of the Vickrey (1969) bottleneck model (but see footnote 2). Equilibrium is characterized for four pricing regimes (in order of increasing sensitivity: no toll, the optimal uniform toll, the optimal step ["coarse"] toll, and the optimal time-varying ["fine"] toll) with capacity exogenous and then with capacity chosen optimally. The four regimes are then compared.

[2]Arnott et al. (1990) formalizes Vickrey's (1969) model, retaining his assumption of inelastic trip demand, and extends it to treat a coarse toll and to solve for optimal capacity. Cohen (1987) and Arnott et al. (1989) extend the Vickrey model to allow for driver heterogeneity. Braid (1989), who wrote his paper independently of ours, compares the social optimum and competitive equilibrium when trip demand is elastic and capacity fixed, with a general trip-cost function. Finally, Arnott and Kraus (1990) examines Ramsey pricing, employing the bottleneck model.

Among other results, we show that, with demand elasticity less than unity, optimal capacity is larger the less sensitive is the (optimal) toll.

Second, the paper provides insight into the appropriate use and interpretation of the standard model of a congestible facility with peak-period congestion. In the Vickrey (1969) model the departure rate from home over the rush-hour period adjusts such that trip price is constant over the period. Equilibrium is therefore determined simultaneously over the entire period, and it makes no sense to compute equilibrium over only part of the period. Thus, the standard procedure of dividing up the period of use into intervals and solving separately for equilibrium in each interval is conceptually unsound. We show, however, that the standard model can be interpreted as providing a reduced-form representation of the structural Vickrey model (since price and the number of users are endogenous, "semi-reduced" form is perhaps more appropriate terminology), as long as the entire rush-hour period is treated as a single interval.

Third, the results of the paper reinforce those in Arnott et al. (1990) in suggesting that the gains from the efficient pricing of road congestion may be considerably greater than those estimated in previous studies (reviewed in Arnott et al. [1990]). Pricing influences an individual's choice of home location, lot size, trip frequency, mode and route choice, and departure time. Also, by influencing land rents, the form and level of urban transport pricing affects urban spatial structure. The literature in urban economics has treated the rush hour as a period of fixed length with uniform congestion. In doing so, it has been able to incorporate all the margins of adjustment noted above except individuals' departure-time decisions. That this paper and Arnott et al. (1990) obtain efficiency gains from road tolling that are substantially greater than those previously estimated suggests that a considerable fraction of the gains result from the change induced in the time pattern of road usage over the peak period. This in turn suggests that more serious consideration should be given to pricing schemes in urban transport that smooth the peak; it also suggests that the benefits from employing technologically sophisticated pricing schemes for urban roads, of the sort advocated for many years by Vickrey (1963, 1971) and discussed in Sanford F. Borins (1986) and Kraus (1989), may have been substantially underestimated.

Fourth, the paper discusses the degree to which roads should be self-financing. Mohring and Harwitz (1962) and Strotz (1965) (MHS hereafter) proved that if the total cost associated with a road (private costs plus capacity cost) doubles when both capacity and usage double, then revenue from the optimal toll equals the cost of constructing and maintaining optimal capacity, so that the road should be self-financing. MHS also examined the cases in which total cost is homogeneous of degree greater and less than one in capacity and usage.[3] We show that the MHS results extend to the basic bottleneck model independently of the tolling regime, as long as the toll is set optimally conditional on the regime. This result has important implications. For example, if toll collection on roads is prohibitively expensive and a gasoline tax is the only policy instrument available to the government which affects the price of a trip, then with optimal capacity (conditional on a gasoline tax) the tax rate should be raised to the point at which revenues from the tax satisfy the MHS conditions.

We shall develop our analysis using the Vickrey (1969) model with identical individuals. Subsequently, we shall discuss briefly how our results extend to heterogeneous individuals. Finally, we shall comment on how the structural approach to modeling urban road congestion can be adapted to develop structural models of other congestible facilities.

Section I examines the economics of the basic Vickrey (1969) model of road congestion. Equilibria with elastic trip demand and fixed capacity for the four pricing regimes are solved for and compared in Section II.

[3]Mohring and Harwitz (1962) and Strotz (1965) developed their models specifically in the context of transport congestion. Williamson (1966) derived the same result in a more specific model, but with a broader context.

The analysis is extended to treat optimal capacity in Section III. Section IV investigates the self-financing results in this context. A numerical example is presented in Section V. Generalizations are discussed in Section VI, and concluding comments are given in Section VII.

I. Review of the Basic Model: Fixed Number of Drivers

A. *The Characteristics of Demand and the Congestion Technology*

Every morning N identical individuals travel from home to work along the same road, each in her own car. Travel is uncongested except at a single bottleneck through which at most s cars can pass per unit time; if the arrival rate at the bottleneck exceeds s, a queue develops. Thus, the capacity constraint is a flow constraint, while the queue discipline is first-come, first-served (FIFO).

Travel time from home to work is

$$(1) \qquad T(t) = T^{f} + T^{v}(t)$$

where T^{f} is fixed travel time, T^{v} is variable travel time, and t is departure time from home. Without affecting results of interest, we set $T^{f} = 0$; thus an individual arrives at the bottleneck as soon as she leaves home and arrives at work immediately upon leaving the bottleneck. Let $D(t)$ be the queue length (i.e., number of cars). Then, an individual's queuing time equals queue length at the time she joins the queue divided by bottleneck capacity:

$$(2) \qquad T^{v}(t) = \frac{D(t)}{s}.$$

Let $\hat{t}$ denote the most recent time at which there was no queue and let $r(t)$ be the departure rate function (from home). Then,

$$(3) \qquad D(t) = \int_{\hat{t}}^{t} r(u)\,du - s(t - \hat{t}).$$

The price of a trip is the sum of private cost and the toll:

$$(4) \qquad p(t) = C(t) + \mathcal{T}(t).$$

Private cost, in turn, is taken to be linear in travel time and schedule delay (time early or time late):

$$(5) \qquad C(t) = \alpha T^{v}(t) + \beta(\text{time early}) + \gamma(\text{time late})$$

where α is the shadow cost of travel time, β is the unit cost of arriving early at work, and γ is the unit cost of arriving late. In accordance with empirical results (Small, 1982), we assume that $\gamma > \alpha > \beta$. Following convention, we call the cost of arriving at work early or late "schedule delay cost" and call $T^{v}(t)$ "travel time cost." The desired arrival time is t^{*}. Thus, "time early" is $\max[0, t^{*} - t - T^{v}(t)]$ and "time late" is $\max[0, t + T^{v}(t) - t^{*}]$.

Each individual decides when to leave home. In doing so, she trades off travel time, schedule delay, and the toll. Equilibrium obtains when no individual can decrease her trip price by altering her departure time, taking all other drivers' departure times as fixed. Thus, the equilibrium is a pure-strategy Nash equilibrium with departure times as the strategy variables. This seems to be the natural equilibrium concept to employ in this context.[4]

B. *The No-Toll Equilibrium*

Let t_{q} be the beginning of the rush hour, let $t_{q'}$ be the end, and let $\tilde{t}$ be the departure time for on-time arrival $[\tilde{t} = t^{*} - T^{v}(\tilde{t})]$. For early arrival $[t \in [t_{q}, \tilde{t})]$, the equal-trip-price condition is

$$(6) \qquad p(t) = \bar{p} = \alpha T^{v}(t) + \beta[t^{*} - t - T^{v}(t)].$$

[4] In this context, pure strategies appear to be more realistic than mixed strategies, since most individuals prefer a routine—to leave home at the same time every day. The mixed-strategy equilibrium is the same if aggregate stochasticity is ignored. Moshe Ben Akiva et al. (1986) demonstrate convergence to the no-toll Nash equilibrium via a specific adjustment process in which, out of equilibrium, individuals adjust their departure times according to a particular probabilistic decision rule.

Differentiation of (6) yields

$$\frac{dT^v(t)}{dt} = \frac{\beta}{\alpha - \beta}. \tag{7}$$

From (3),

$$\frac{dD(t)}{dt} = r(t) - s. \tag{8}$$

Combining (2), (7), and (8) gives

$$r(t) = \frac{\alpha s}{\alpha - \beta} \qquad \text{for } t \in [t_q, \tilde{t}). \tag{9}$$

By a similar argument, it can be shown that

$$r(t) = \frac{\alpha s}{\alpha + \gamma} \qquad \text{for } t \in [\tilde{t}, t_{q'}]. \tag{10}$$

Thus, the queue length evolves over the rush hour such that the equal-trip-price condition is satisfied.

The individuals who depart at the beginning and end of the rush hour incur only schedule delay cost, which must be equal in equilibrium. Since arrivals are continuous over the rush hour, the length of the rush hour is N/s. These results together imply

$$t_q = t^* - \left(\frac{\gamma}{\beta + \gamma}\right)\frac{N}{s} \tag{11a}$$

$$t_{q'} = t^* + \left(\frac{\beta}{\beta + \gamma}\right)\frac{N}{s}. \tag{11b}$$

Furthermore, it is easily shown that

$$\tilde{t} = t^* - \left(\frac{\beta}{\alpha}\right)\left(\frac{\gamma}{\beta + \gamma}\right)\frac{N}{s}. \tag{12}$$

The solution is depicted in Figure 1. The vertical distance between the cumulative departures schedule and the cumulative arrivals schedule is queue length, and the horizontal distance is travel time [$D(t')$ and $T^v(t')$, respectively, in the figure]. The queue builds up linearly from t_q to $\tilde{t}$, and then dissipates linearly until it disappears at $t_{q'}$.

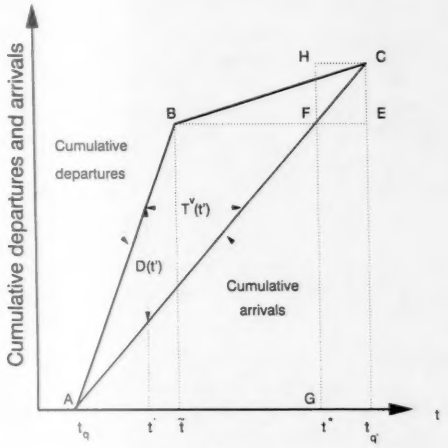


FIGURE 1. THE NO-TOLL EQUILIBRIUM

Let TTC be total travel time cost, let SDC be total schedule delay cost, and let TC be total travel cost. In Figure 1, TTC is α times area ABCA, and SDC is β times area AFGA plus γ times area CFHC. It is straightforward to show that

$$\mathrm{TTC}^e(N,s) = \mathrm{SDC}^e(N,s) = \frac{\delta}{2}\left(\frac{N^2}{s}\right) \tag{13}$$

$$\mathrm{TC}^e = \delta\left(\frac{N^2}{s}\right)$$

where

$$\delta \equiv \frac{\beta\gamma}{\beta + \gamma} \tag{14}$$

and superscript e denotes the no-toll equilibrium.[5]

Since there is no toll, price equals private cost, which is the same for everyone and equals average total travel cost, ATC ≡

[5]Note that all these aggregates are independent of α. The start and end of the rush hour are independent of α, and therefore, so is total schedule delay cost. Since $\mathrm{TTC}^e = \mathrm{SDC}^e$, TTC^e and TC^e too are independent of α. A doubling of α halves queue length, preserving travel time cost.

TC/N; that is,

$$p^e = C^e = ATC^e. \tag{15}$$

We may calculate marginal social cost (MSC) from (13):

$$MSC^e = \frac{\partial TC^e}{\partial N} = \frac{2\delta N}{s} \tag{16}$$

$$= 2\frac{TC^e}{N} = 2ATC^e.$$

Note that marginal social cost is independent of departure time. The reason is that other drivers adjust their departure times in such a way that the departure-rate function (including the additional traveler) is independent of the additional traveler's departure time.

Several observations are in order:

1. In the no-toll equilibrium, total schedule delay cost equals total travel time cost. The model highlights the costs of traveling at inconvenient times, which are hidden in the standard model.
2. Equation (15), after substitution of (13) and (14), characterizes price as a function of N and s. The resulting function is therefore the reduced-form supply function, to which a trip demand function may be added to determine the full equilibrium. The supply function captures the technology, the pricing regime considered, and commuters' time-of-use decisions. The demand function captures commuters' frequency-of-use decisions.
3. Downs's law (Anthony Downs, 1962), which is based on observation, states that an expansion of road capacity will reduce the length of the peak period more than the level of congestion at the peak. In the bottleneck model with inelastic demand, a doubling of capacity halves the length of the rush hour and the maximum travel time but leaves the maximum queue length unaltered.
4. Queuing time in the model is pure deadweight loss. If the departure rate were set at s between t_q and $t_{q'}$, queuing would be eliminated. Furthermore, since the time pattern of arrivals would be the same as in the no-toll equilibrium, total schedule delay cost would be unchanged. Thus, the social saving from the change would equal total travel time cost in the no-toll equilibrium, which equals one-half of total travel cost. This example provides a stark illustration of the potential social saving that can be achieved through altering the time pattern of departures, even when the total number of trips and the distribution of arrival times are held constant.
5. Private cost and marginal social cost are determined by the pattern of congestion over the entire period of use, and not, as is assumed in the standard analysis, by the number of users over some portion of the period.
6. Marginal social cost should be computed mutatis mutandis, not ceteris paribus; that is, marginal social cost is computed incorrectly if one adds a commuter and computes the increase in social cost from doing so, without allowing other drivers to adjust their departure times. The reason why computing marginal social cost ceteris paribus is incorrect is that the envelope theorem does not hold. Because there is no toll, prices are distorted. Consequently, the adjustments that commuters make in their departure times in response to the added driver alter the deadweight loss from unpriced congestion.
7. We shall see later that a change in the form of pricing alters the private and marginal social cost functions. Thus, contrary to the standard model, the private and marginal social cost functions are not completely technologically determined, but depend as well on the form of pricing. This is because the cost functions capture consumers' time-of-use decisions, which depend on the form of pricing, even with N and s fixed.

II. Elastic Demand, Capacity Arbitrary

In the previous section, we characterized the reduced-form supply function in the ab-

sence of a toll. We now incorporate variable trip demand.

The demand for trips is assumed to be

$$(17) \qquad N = N(p) \qquad \frac{dN}{dp} < 0.$$

The consumers' surplus from travel with price p is

$$(18) \qquad \mathrm{CS}(p) = \int_p^\infty N(p')\,dp'$$

while the social surplus (gross of capacity costs) from travel, with price p and average toll τ, is the sum of consumers' surplus and toll revenues:

$$(19) \qquad \mathrm{SS}(p,\tau) = \mathrm{CS}(p) + N\tau.$$

A. *The No-Toll Equilibrium*

From (13) and (15),

$$(20) \qquad p^e = \frac{\delta N^e}{s}$$

which is the supply function for trips in the absence of a toll. Solving this simultaneously with the demand function (17) yields the equilibrium trip price and number of trips, $\hat{p}^e(s)$ and $\hat{N}^e(s)$, where a "^" over a variable indicates that it is a function of only s (plus exogenous parameters). We also have

$$(21) \qquad \widehat{\mathrm{CS}}^e(s) = \widehat{\mathrm{SS}}^e(s) = \int_{\hat{p}^e(s)}^\infty N(p')\,dp'.$$

We now derive the reduced-form supply functions for the three toll regimes, as was done in the previous section for the case of no toll. Subsequently, we shall characterize equilibrium for the various toll regimes and then compare the four equilibria.

B. *Alternative Tolling Regimes*

We shall consider three different tolling regimes: a uniform toll, a fine toll, and a coarse toll. The *uniform toll* is constant throughout the day. It does not alter the pattern of congestion over the rush hour, but with elastic trip demand it causes a reduction in the number of trips. The *fine toll* is a completely flexible time-dependent toll. The *coarse toll* is intermediate between the uniform and fine tolls. It combines a lower off-peak toll with a higher peak toll and is characterized by four parameters: the magnitudes of the peak and off-peak tolls, as well as the times at which the peak toll is applied and later removed. The average level of the coarse toll discourages travel, and the toll differential between the peak and off-peak periods shifts traffic from the peak to the off-peak period.

For each toll, we assume throughout the paper that the parameters of the toll are set optimally. The derivations of the optimal tolls are presented in Arnott et al. (1990); here we record the results and provide some intuition for them.

Uniform Toll.—Since the uniform toll adds a constant fee to each trip, for fixed N it does not alter the departure pattern [recall the derivations of (9) and (10)]. Thus, total travel cost is related to N and s in the same way as in the no-toll equilibrium; that is,

$$(22) \qquad \mathrm{TC}^u = \delta\left(\frac{N^2}{s}\right)$$

where superscript u denotes the uniform toll. Thus,

$$(23) \qquad \mathrm{MSC}^u = 2\mathrm{ATC}^u.$$

Since price and the toll are the same for all commuters, so too is private cost. Also, average travel cost equals (average) private cost. Thus,

$$(24) \qquad p^u = C^u + \tau^u = \mathrm{ATC}^u + \tau^u.$$

For efficiency, the price of a trip equals marginal social cost. Hence, using (23) and

(24), one obtains

$$(25)\quad \mathrm{ATC}^{\mathrm{u}} + \tau^{\mathrm{u}} = p^{\mathrm{u}} = \mathrm{MSC}^{\mathrm{u}} = 2\mathrm{ATC}^{\mathrm{u}}$$

$$\tau^{\mathrm{u}} = \mathrm{ATC}^{\mathrm{u}}.$$

Thus, the optimal toll equals average travel cost.

Fine Toll.—The social optimum can be decentralized with a time-varying toll. The social optimum entails no queue and therefore zero total travel time cost. Also, for a given number of travelers, the beginning and end of the rush hour are the same as in the no-toll equilibrium. Thus, $\mathrm{SDC}^{\mathrm{o}}(N,s) = \mathrm{SDC}^{\mathrm{e}}(N,s)$ while $\mathrm{TTC}^{\mathrm{o}}(N,s) = 0$, so that, from (13),

$$(26)\qquad \mathrm{TC}^{\mathrm{o}} = \mathrm{SDC}^{\mathrm{o}} = \frac{\delta}{2}\left(\frac{N^2}{s}\right)$$

where a superscript o denotes the social optimum. Efficiency requires that each traveler pays marginal social cost which equals twice the average travel cost. In the absence of a queue the person who arrives on time faces zero travel cost and for efficiency must therefore pay a toll equal to twice the average travel cost. Furthermore, the persons who depart first and last incur twice the average schedule delay and hence twice the average travel cost and should therefore pay no toll. Finally, because of the linearity of (5), the fine toll increases linearly from t_{q} to t^* and then decreases linearly from t^* to $t_{\mathrm{q}'}$. The average toll equals average travel cost.

Coarse Toll.—For this toll, too, the average toll paid equals average travel cost. This toll is more efficient than a uniform toll since it alters the queuing pattern over the rush hour, but since it does not completely eliminate queuing, it is not as efficient as the fine toll. It is shown in Arnott et al. (1990) that, with application of the coarse toll,

$$(27)\quad \mathrm{TC}^{\mathrm{c}} = \frac{\delta}{4}\left[3 - \frac{(\gamma-\alpha)\beta}{(\beta+\gamma)(\alpha+\beta)}\right]\frac{N^2}{s}$$

where superscript c denotes the coarse toll.[6] We collect the above results in the following proposition.

PROPOSITION 1: *Let* $j = \mathrm{e,u,c,o}$ *index the pricing regime. Then,*

$$(28\mathrm{a})\qquad \mathrm{TC}^j(N,s) = \frac{\Gamma^j \delta N^2}{s}$$

$$(28\mathrm{b})\quad \mathrm{ATC}^j(N,s) = \frac{\Gamma^j \delta N}{s}$$

$$(28\mathrm{c})\quad \mathrm{MSC}^j(N,s) = 2\mathrm{ATC}^j(N,s)$$

where

$$(29\mathrm{a})\quad \Gamma^{\mathrm{e}} = \Gamma^{\mathrm{u}} = 1$$

$$(29\mathrm{b})\quad \Gamma^{\mathrm{o}} = \tfrac{1}{2}$$

$$(29\mathrm{c})\quad \Gamma^{\mathrm{c}} = \frac{1}{4}\left[3 - \frac{(\gamma-\alpha)\beta}{(\beta+\gamma)(\alpha+\gamma)}\right].$$

Also,

$$(30\mathrm{a})\qquad p^{\mathrm{e}}(N,s) = \mathrm{ATC}^{\mathrm{e}}(N,s).$$

Furthermore, for $j = \mathrm{u,c,o}$, *where* τ^j *denotes the corresponding average toll,*

$$(30\mathrm{b})\quad p^j(N,s) = \mathrm{ATC}^j(N,s) + \tau^j(N,s)$$

and

$$(31)\qquad p^j(N,s) = \mathrm{MSC}^j(N,s) = 2\mathrm{ATC}^j(N,s)$$

[6]The (optimal) coarse toll has the following characteristics. It is applied at the front of the queue. The peak toll is turned on before t^* and off after t^*. In the early morning, well before the peak toll is applied, a queue builds up. Then, for a period of time immediately before the peak toll is applied, there are no departures, and the queue length decreases to zero at the moment the peak toll takes effect. The queue then rises until t^* and falls after t^* until it reaches zero just before the peak toll is lifted. Right after the peak toll is lifted, there is a mass of departures and no departures subsequently.

which together imply

$$(32)\quad \tau^j(N,s) = \mathrm{ATC}^j(N,s) = \frac{\mathrm{MSC}^j(N,s)}{2} = \frac{p^j(N,s)}{2}.$$

With the tolls set at their optimal levels, commuters pay the marginal social cost of each trip. Thus, the marginal-social-cost functions for the three toll regimes given in (28) and (29) are the corresponding reduced-form supply functions.

C. *Equilibrium for the Toll Regimes*

From (28) and (31), for $j = \mathrm{u,c,o}$, the trip supply function for the various toll regimes is

$$(33)\qquad p^j(N,s) = \frac{2\Gamma^j \delta N}{s}$$

and equations (33) and (17) together implicitly give $\hat{N}^j(s)$ and $\hat{p}^j(s)$, while (32), (33), and (17) give $\hat{\tau}^j(s)$. Equilibrium for a particular toll regime is portrayed graphically in Figure 2. The diagram is standard. However, the demand and cost curves are defined over the entire rush hour, and it makes no sense to define them over an interval of the rush hour. Thus, the standard graphical treatment of traffic congestion is consistent with the analysis of this paper when the rush hour is treated as a single period, but not when the rush hour is divided into time intervals.

We also have, for $j = \mathrm{u,c,o}$,

$$(34)\qquad \widehat{\mathrm{CS}}^j(s) = \int_{\hat{p}^j(s)}^{\infty} N(p)\,dp$$

and

$$(35)\quad \widehat{\mathrm{SS}}^j(s) = \widehat{\mathrm{CS}}^j(s) + \hat{N}^j(s)\hat{\tau}^j(s).$$

D. *Comparison of the Four Regimes*

Together (28), (29), (31), and (17) imply

$$(36)\quad \hat{p}^{\mathrm{u}}(s) > \hat{p}^{\mathrm{c}}(s) > \hat{p}^{\mathrm{o}}(s) = \hat{p}^{\mathrm{e}}(s)$$

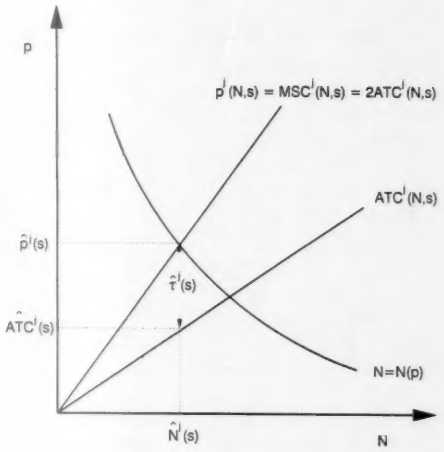


FIGURE 2. EQUILIBRIUM WITH THE OPTIMAL TOLL UNDER TOLL REGIME j: $\hat{\tau}^j(s) = \widehat{\mathrm{ATC}}^j(s)$

which, with variable demand, in turn implies

$$(37)\quad \hat{N}^{\mathrm{u}}(s) < \hat{N}^{\mathrm{c}}(s) < \hat{N}^{\mathrm{o}}(s) = \hat{N}^{\mathrm{e}}(s).$$

The explanation of these inequalities is straightforward. For a given number of travelers, the price of a trip is lowest for the no-toll equilibrium (since no toll is charged) and the fine-toll equilibrium (since travel on the road is efficient), intermediate for the coarse-toll equilibrium, and highest for the uniform-toll equilibrium. The coincidence of the no-toll and fine-toll supply functions reflects the particular congestion technology assumed.[7]

One can calculate the efficiency loss, relative to the fine toll, of the other pricing regimes:

$$(38)\quad \widehat{\mathrm{EL}}^j(s) = \widehat{\mathrm{SS}}^{\mathrm{o}}(s) - \widehat{\mathrm{SS}}^j(s) \qquad j = \mathrm{c,u,e}.$$

[7] In particular, $\mathrm{SDC}^{\mathrm{e}} = \mathrm{TTC}^{\mathrm{e}} = \mathrm{SDC}^{\mathrm{o}}$ and $\mathrm{TTC}^{\mathrm{o}} = 0$, which imply that $\mathrm{ATC}^{\mathrm{e}} = 2\mathrm{ATC}^{\mathrm{o}}$, so that $p^{\mathrm{o}} = 2\mathrm{ATC}^{\mathrm{o}} = \mathrm{ATC}^{\mathrm{e}} = p^{\mathrm{e}}$.

The efficiency loss in the no-toll equilibrium equals the toll revenue collected at the fine-toll equilibrium. The efficiency loss in the uniform-toll and coarse-toll equilibria equals the corresponding loss in consumers' surplus plus the loss (possibly negative) of toll revenue. Since the no-toll situation is a special case of a uniform toll, which is in turn a special case of a coarse toll, it is evident that

$$(39)\quad \widehat{\mathrm{EL}}^{\mathrm{e}}(s) > \widehat{\mathrm{EL}}^{\mathrm{u}}(s) > \widehat{\mathrm{EL}}^{\mathrm{c}}(s) > 0.$$

III. Elastic Demand, Optimal Capacity

Optimal capacity is that capacity which maximizes gross social surplus less capacity costs; it occurs where the marginal benefit from capacity expansion (the gain in social surplus) equals the marginal cost. Let $K(s)$ be the capacity–construction-cost function. The marginal capacity-expansion cost is

$$(40)\qquad \mathrm{MC}(s) = \frac{dK(s)}{ds}.$$

A. *The No-Toll Regime*

Since there is no toll, the marginal benefit is simply the gain in consumers' surplus. From (21),

$$(41)\quad \widehat{\mathrm{MB}}^{\mathrm{e}}(s) = \frac{d\widehat{\mathrm{CS}}^{\mathrm{e}}(s)}{ds} = -\hat{N}^{\mathrm{e}}(s)\frac{d\hat{p}^{\mathrm{e}}(s)}{ds}.$$

B. *The Toll Regimes* ($j = \mathrm{u,c,o}$)

From (35), the marginal benefit from capacity expansion equals the marginal consumers' surplus plus the marginal toll revenue; that is,

$$(42)\quad \widehat{\mathrm{MB}}^{j}(s) = \frac{d\widehat{\mathrm{CS}}^{j}(s)}{ds} + \frac{d[\hat{N}^{j}(s)\hat{\tau}^{j}(s)]}{ds}.$$

From (34),

$$\frac{d\widehat{\mathrm{CS}}^{j}(s)}{ds} = -\hat{N}^{j}(s)\frac{d\hat{p}^{j}(s)}{ds}$$

and from (32),

$$\frac{d\hat{\tau}^{j}(s)}{ds} = \left(\frac{1}{2}\right)\frac{d\hat{p}^{j}(s)}{ds}.$$

Thus, from the above, and using (28b) and (32),

$$\begin{aligned}(43)\quad \widehat{\mathrm{MB}}^{j}(s) &= -\hat{N}^{j}(s)\frac{d\hat{p}^{j}(s)}{ds} + \frac{d\hat{N}^{j}(s)}{ds}\hat{\tau}^{j}(s) + \hat{N}^{j}(s)\frac{d\hat{\tau}^{j}(s)}{ds}\\ &= \frac{d\hat{N}^{j}(s)}{ds}\hat{\tau}^{j}(s) - \hat{N}^{j}(s)\frac{d\hat{\tau}^{j}(s)}{ds}\\ &= \hat{N}^{j}(s)\hat{\tau}^{j}(s)\left[\frac{1}{\hat{N}^{j}(s)}\frac{d\hat{N}^{j}(s)}{ds} - \frac{1}{\hat{\tau}^{j}(s)}\frac{d\hat{\tau}^{j}(s)}{ds}\right]\\ &= \frac{\hat{N}^{j}(s)\hat{\tau}^{j}(s)}{s}\\ &= \frac{\widehat{\mathrm{TC}}^{j}(s)}{s}\end{aligned}$$

which has the following interpretation: since the tolls are set at the optimal level conditional on the tolling regime (with the result that commuters pay the marginal social cost of a trip), the analysis is first-best.[8] As a result, the envelope theorem holds, which in this context implies that the marginal benefit of an incremental capacity expansion is the same before and after individuals adjust

[8]There is a difficulty of terminology here. There are two qualitatively different sources of inefficiency possible: the *level* (or absence) and the *form* of the toll. We refer to a situation where travelers face the marginal social cost of a trip as first-best, even when the form of the toll is not optimal.

trip frequency in response to the capacity expansion. Since total travel cost with N fixed is proportional to the inverse of s, the marginal benefit of capacity expansion with trip frequency fixed is $\widehat{\mathrm{TC}}^j/s$, which equals the marginal benefit of capacity expansion with trip frequency variable.

Recall from (36) that $\hat{p}^{\mathrm{u}}(s) > \hat{p}^{\mathrm{c}}(s) > \hat{p}^{\mathrm{o}}(s)$. With elasticity of demand (ε) less than unity, it follows that $\hat{N}^{\mathrm{u}}(s)\hat{p}^{\mathrm{u}}(s) > \hat{N}^{\mathrm{c}}(s)\hat{p}^{\mathrm{c}}(s) > \hat{N}^{\mathrm{o}}(s)\hat{p}^{\mathrm{o}}(s)$, which in turn implies from (43) and (31) that

$$(44)\quad \widehat{\mathrm{MB}}^{\mathrm{u}}(s) > \widehat{\mathrm{MB}}^{\mathrm{c}}(s) > \widehat{\mathrm{MB}}^{\mathrm{o}}(s) \qquad \text{for } \varepsilon < 1.$$

We assume that for each regime the marginal-cost curve cuts the marginal-benefit curve from below. Then optimal capacity is determined by setting marginal benefit equal to marginal cost, and from (40) and (44)

$$(45)\quad s_*^{\mathrm{u}} > s_*^{\mathrm{c}} > s_*^{\mathrm{c}} \qquad \text{for } \varepsilon < 1$$

where "$*$" subscripts denote "at optimal capacity." Thus, when $\varepsilon < 1$, optimal capacity is larger the less sensitive the tolling regime. The less sensitive the tolling regime, the higher is average travel cost given capacity. Furthermore, since price equals twice average travel cost for each toll regime, with $\varepsilon < 1$ total travel cost is higher the coarser is the tolling regime given capacity. Finally, the marginal benefit from capacity expansion, for each s, is directly proportional to total travel cost. Hence, for each level of capacity, the marginal benefit from capacity expansion is higher, the less sensitive the tolling regime. The same line of argument implies that when $\varepsilon > 1$, optimal capacity is smaller, the coarser the toll regime.

From (40) and (43) and using $p_*^j = 2\mathrm{ATC}_*^j = 2\mathrm{TC}_*^j/N_*^j$,

$$N_*^j p_*^j = 2s_*^j \left.\frac{dK(s)}{ds}\right|_{s_*^j}.$$

Thus, as long as the elasticity of marginal construction costs with respect to capacity is greater than -1, a condition we assume to hold,

$$(46)\quad N_*^{\mathrm{u}} p_*^{\mathrm{u}} > N_*^{\mathrm{c}} p_*^{\mathrm{c}} > N_*^{\mathrm{o}} p_*^{\mathrm{o}} \qquad \text{for } \varepsilon < 1.$$

This implies that, for $\varepsilon < 1$,

$$(47)\quad p_*^{\mathrm{u}} > p_*^{\mathrm{c}} > p_*^{\mathrm{o}}$$

and, with $\varepsilon > 0$,

$$N_*^{\mathrm{u}} < N_*^{\mathrm{c}} < N_*^{\mathrm{o}}.$$

The cases of $\varepsilon = 1$ and $\varepsilon > 1$ can be found in Arnott et al. (1987).[9]

C. *Comparison of No-Toll and Toll Regimes*

It remains to compare the no-toll equilibrium with optimal capacity to the various toll equilibria with optimal capacity. Unfortunately, the results are not as clean as for the comparison of the toll equilibria with optimal capacity since, with the price of a trip in the no-toll equilibrium below marginal social cost, the determination of optimal capacity is a second-best problem. Equation (43) does not apply in the no-toll equilibrium because the envelope theorem does not hold (the marginal benefit from capacity expansion in the no-toll equilibrium includes the change in the efficiency loss due to underpriced congestion caused by the induced increase in trip frequency). However, some results can be obtained, and these are presented in Arnott et al. (1987).

The empirical evidence strongly supports the hypothesis that rush-hour car travel, which is predominantly commuting, is price-inelastic (e.g., Daniel McFadden, 1974; Small, 1983; John Pucher and Jerome Rothenberg, 1976). Combining this stylized fact with (45), we have the prediction that optimal capacity falls as the "sensitivity of the tolling regime" increases. The model also predicts that, with constant elasticity of

[9] It is interesting to note that, with optimal capacity and constant unit construction cost, $\widehat{\mathrm{MB}}^j(s) = \widehat{\mathrm{TC}}^j/s = k \equiv \mathrm{MC}(s)$, which implies $\mathrm{TC}_*^j = ks_*^j$; that is, capacity construction cost equals total travel cost. This result is due to the form of congestion.

demand less than 1, optimal capacity in the no-toll regime is higher than under all the tolling regimes (see Arnott et al., 1987).

It is worth restating that there are two separate issues involved in comparing optimal capacity between regimes. The first is whether "second-best capacity" (optimal capacity conditional on congestion being underpriced) is less than or greater than first-best capacity (where congestion is efficiently priced). This is a classic problem that has been well researched in the context of the standard model of traffic congestion (e.g., William C. Wheaton, 1978; John D. Wilson, 1983; Edmond L. d'Ouville and John F. McDonald, 1990). Broadly speaking, our results are consistent with these papers, though not fully comparable because their treatment of congestion is static. The second issue entails the comparison of first-best toll regimes that differ in their sensitivity. Here our analysis is new, and the results are clean.

For future reference, we note that with optimal capacity the efficiency losses associated with the coarse-, uniform-, and no-toll equilibria compared to the optimal fine-toll equilibrium are

$$(48)\quad \mathrm{EL}^j_* = \begin{cases} \displaystyle\int_{p^o_*}^{p^e_*} N(p)\,dp + \frac{N^o_* p^o_*}{2} - K(s^o_*) + K(s^e_*) & j = \mathrm{e} \\ \displaystyle\int_{p^o_*}^{p^j_*} N(p)\,dp + \frac{N^o_* p^o_*}{2} - \frac{N^j_* p^j_*}{2} - K(s^o_*) + K(s^j_*) & j = \mathrm{u,c.} \end{cases}$$

The results of the previous two sections are brought together in the following proposition.

PROPOSITION 2:

(i) $\hat{p}^{\mathrm{u}}(s) > \hat{p}^{\mathrm{c}}(s) > \hat{p}^{\mathrm{o}}(s) = \hat{p}^{\mathrm{e}}(s)$;

(ii) $\widehat{\mathrm{EL}}^{\mathrm{e}}(s) > \widehat{\mathrm{EL}}^{\mathrm{u}}(s) > \widehat{\mathrm{EL}}^{\mathrm{c}}(s) > 0$, $\mathrm{EL}^{\mathrm{e}}_* > \mathrm{EL}^{\mathrm{u}}_* > \mathrm{EL}^{\mathrm{c}}_*$;

(iii) *with price-sensitive trip demand,* $\hat{N}^{\mathrm{u}}(s) < \hat{N}^{\mathrm{c}}(s) < \hat{N}^{\mathrm{o}}(s) = \hat{N}^{\mathrm{e}}(s)$;

(iv) *with* $\varepsilon < 1$, $s^{\mathrm{u}}_* > s^{\mathrm{c}}_* > s^{\mathrm{o}}_*$;

(v) *with* $\varepsilon < 1$ *and* $[d^2K(s)/ds^2]s + dK(s)/ds > 0$, $p^{\mathrm{u}}_* > p^{\mathrm{c}}_* > p^{\mathrm{o}}_*$, *and (if* $\varepsilon > 0$*)* $N^{\mathrm{u}}_* < N^{\mathrm{c}}_* < N^{\mathrm{o}}_*$.

IV. The Self-Financing of Capacity

Mohring and Harwitz (1962) and Strotz (1965) investigated the extent to which a network of roads should be self-financing. In the single-period version of their analysis, the private cost (C) is a function of capacity (s) and the number of trips (N). The planner chooses price and capacity so as to maximize social surplus [consumers' surplus plus toll revenue, $R = (p - C)N$] minus capacity costs, $K(s)$; that is,

$$(49)\quad \max_{p,s} \int_p^{\infty} N(p')\,dp' + [p - C(N(p), s)]N(p) - K(s).$$

The corresponding first-order conditions for p and s are:

$$(50\mathrm{a})\quad -N + \left(1 - \frac{\partial C}{\partial N}\frac{dN}{dp}\right)N + (p - C)\frac{dN}{dp} = 0$$

and

$$(50\mathrm{b})\quad -\frac{\partial C}{\partial s}N - \frac{dK}{ds} = 0$$

respectively. Equation (50a) reduces to

$$(50\mathrm{a}')\quad p = C + \frac{\partial C}{\partial N}N.$$

Equation (50a′) states that price should be set equal to marginal social cost, the sum of the private cost (C) and the marginal congestion externality ($[\partial C/\partial N]N$), and (50b) states that capacity should be constructed up to the point where marginal cost (dK/ds) equals marginal benefit ($-[\partial C/\partial s]N$). Assume to simplify that $C(\cdot)$ is homogeneous of degree h^{c} in N and s and that $K(\cdot)$ is homogeneous of degree h^{K} in s. Then, using Euler's Theorem, (50b)

may be rewritten as

$$(51) \qquad N\left(\frac{\partial C}{\partial N}N\right) = h^{K}K + Nh^{C}C.$$

The optimal toll, τ, should be set equal to the marginal congestion externality, $(\partial C/\partial N)N$. Thus,

$$(52) \qquad R \equiv N\tau = h^{K}K + Nh^{C}C$$

where R is toll revenue. Equation (52) implies that, if there are constant costs to capacity expansion so that $h^{K} = 1$ and if a doubling of capacity and the number of trips leaves the cost of a trip unchanged so that $h^{c} = 0$, then $R = K$; that is, optimal toll revenue exactly covers the cost of constructing optimal capacity. One can extend the analysis to compute the proportion of the cost of capacity construction that is financed from toll revenue when either the toll or the level of capacity is nonoptimal.

The issue of interest here is the form of the self-financing results in the bottleneck model. Obviously, in the no-toll regime, no toll revenue is collected. We therefore restrict our attention to the uniform, coarse, and fine tolls. The analog to (49) is

$$(53) \qquad \max_{p^j, s} \int_{p^j}^{\infty} N(p')\,dp' + \left[p^j - \mathrm{ATC}^j\left(N(p^j), s\right)\right]N(p^j) - K(s) \qquad j = \mathrm{u,c,o}.$$

It is evident that (53) is simply a particularization of (49) and, hence, that the MHS self-financing results hold for our model. Since $\mathrm{ATC}(\cdot)$ is homogeneous of degree zero in N and s for each tolling regime (recall Proposition 1), equation (52) reduces to

$$(54) \qquad R^j_* = h^{K}K\left(s^j_*\right) \qquad j = \mathrm{u,c,o}.$$

Independent of the tolling regime, the ratio of the revenue collected from the optimal toll to the construction cost of optimal capacity equals the elasticity of construction cost with respect to capacity. This is remarkable since it indicates that *the optimal degree of self-financing of a road is independent of the form of the pricing system employed*; for example, if a road system should be self-financing when a sophisticated tolling system is employed, it should also be self-financing when only a flat parking fee is applied.

The major result of this section is summarized in the following proposition.

PROPOSITION 3: *The ratio of the revenue collected from the optimal toll to the costs of constructing optimal capacity equals the elasticity of construction cost with respect to capacity, whatever the tolling regime. With constant costs of capacity, the road should be self-financing; with increasing costs, it should generate a surplus; and with decreasing costs, it should operate at a loss.*

V. A Numerical Example

We start by considering the situation in which capacity is fixed. Subsequently, we extend the example to treat optimal capacity.

A. *Capacity Fixed*

We normalize so that $N^{e} = 1$, and then set $s = 0.4$ cars/hr so that the length of the rush hour in the no-toll equilibrium, $t_{q'} - t_{q} = N^{e}/s = 2.5$ hours. Thus, $p^{e} = \delta(N^{e}/s)$ [from (20)] $= \$2.5(\delta)$. Consistent with column 1 of table 2 in Small (1982), we take $\alpha = \$5.00/\mathrm{hr}$, $\beta = \$3.05/\mathrm{hr}$, and $\gamma = \$11.88/\mathrm{hr}$.[10] Hence, $\delta \equiv \beta\gamma/(\beta + \gamma) = \$2.425/\mathrm{hr}$, $p^{e} = \$6.063$, and $\Gamma^{c} = 0.7292$

[10]Small (1982) estimated the ratios β/α and γ/α. To construct an estimate of α we draw on Small (1991 pp. 2–54), who concludes that "...a reasonable average value of time for journey to work is 50 percent of the gross wage rate...." Bureau of Labor Statistics (1991) reports average 1990 hourly earnings of \$10.03. Half this is \$5.00/hr in round figures, which we take as our value of α. Since the Bureau of Labor Statistics figure excludes supervisory and government workers, it understates the average wage for the whole population and thus leads to a conservative value for α as constructed.

TABLE 1—NUMBER OF COMMUTERS, COST OF TRIP, AND EFFICIENCY LOSS WITH VARIOUS TOLLING REGIMES WHEN ROAD CAPACITY IS FIXED ($N^c = 1$, $s = 0.4$ vehicles/hr, $\alpha = \$5.00$/hr, $\beta = \$3.05$/hr, $\gamma = \$11.88$/hr)

Variable	ε	Tolling regime e	u	c	o
$\hat{N}^j$	0	1	1	1	1
	0.2	1	0.8909	0.9390	1
	1	1	0.7071	0.8280	1
$\hat{p}^j$	0	6.063	12.125	8.842	6.063
	0.2	6.063	10.802	8.304	6.063
	1	6.063	8.574	7.322	6.063
$\widehat{\mathrm{EL}}^j$	0	3.031	3.031	1.390	0
	0.2	3.031	2.671	1.302	0
	1	3.031	2.101	1.144	0
$\left(\frac{\widehat{\mathrm{EL}}^j}{\widehat{\mathrm{EL}}^c}\right)$	0	1	1	0.458	0
	0.2	1	0.881	0.429	0
	1	1	0.693	0.377	0

[from (29c)]. Finally, we assume a constant-elasticity demand function, $N = np^{-\varepsilon}$. Then $n = N^c(p^c)^\varepsilon = 1$ for $\varepsilon = 0$, 1.434 for $\varepsilon = 0.2$, and 6.063 for $\varepsilon = 1$. Throughout the example, we treat three demand elasticities: (i) $\varepsilon = 0$, the extreme of completely inelastic trip demand; (ii) $\varepsilon = 1$, the highest possible reasonable value; and (iii) $\varepsilon = 0.2$, our best guess on the basis of McFadden (1974), Pucher and Rothenberg (1976), and Small (1983).[11]

From formulas in Arnott et al. (1987), we obtain the results given in Table 1. Recall that the average toll equals one-half the price in the three tolling regimes. Several points are worthy of note. First, observe that the uniform toll tends to be more efficient (i.e., $\widehat{\mathrm{EL}}^u/\widehat{\mathrm{EL}}^c$ is lower) the higher the elasticity of demand. This is as expected. There are two sources of efficiency loss from imposing no toll: first, because travel is underpriced, too many trips are taken; second, given the number of trips taken, cars do not distribute themselves efficiently over the rush hour. The uniform toll affects only the first source of efficiency loss, which is relatively more important the more elastic is trip demand. Second, with our best-guess demand elasticity of 0.2, more than half the efficiency gain from applying the fine toll can be achieved by applying the coarse toll. Third, the efficiency gain from applying congestion tolls can be substantial: relative to the no-toll situation, \$3.03 per trip can be saved by applying the optimal fine toll. As argued in Arnott et al. (1990), this suggests that the potential gain from applying urban auto tolls may have been considerably underestimated. Fourth, with our best-guess demand elasticity, the efficiency gain from applying the optimal coarse toll is almost five times as high as that from applying the optimal uniform toll. Thus, the efficiency gain from the rescheduling of trips induced by an urban auto toll is likely to be larger than the efficiency gain from the reduction in trips caused by the toll. This in turn strengthens the argument for more serious investigation of technologically advanced tolling schemes.

B. *Capacity Optimal*

We assume that there are constant costs to capacity expansion [$K(s) = ks$] and that construction costs are such that the length of the rush hour is 2.5 hours when road width is chosen optimally for the no-toll regime. Using formulas in Arnott et al. (1987), we find $k = 15.157$ for $\varepsilon = 0$, $k = 12.630$ for $\varepsilon = 0.2$, and $k = 7.578$ for $\varepsilon = 1$, and we obtain the results given in Table 2. They are qualitatively much the same as those presented in Table 1 and are consistent with the results given in Proposition 2. It is worth remarking, however, that with our best-guess demand elasticity, optimal capacity with the fine toll is only about 75 percent of that with no toll.

[11] Our model is sufficiently stylized that it is not clear what elasticity should be employed, though it should be long-run. Note also that since we set fixed travel time to zero, ε is the elasticity of trips with respect to the *variable* component of trip price. The empirical literature, on the other hand, measures the elasticity of trips with respect to the *total* trip price.

TABLE 2—NUMBER OF COMMUTERS, PRICE OF TRIP, AND EFFICIENCY LOSS WITH VARIOUS TOLLING REGIMES WHEN ROAD CAPACITY IS SET OPTIMALLY ($N^c_* = 1$, $s^c_* = 0.4$ vehicles/hr, $\alpha = \$5.00$/hr, $\beta = \$3.05$/hr, $\gamma = \$11.88$/hr)

Variable	ε	Tolling regime			
		e	u	c	o
N^j_*	0	1	1	1	1
	0.2	1	0.8866	0.9150	0.9502
	1	1	0.7071	0.8281	1
p^j_*	0	6.063	12.125	10.355	8.574
	0.2	6.063	11.069	9.452	7.827
	1	6.063	8.574	7.321	6.063
EL^j_*	0	3.551	3.551	1.780	0
	0.2	3.334	2.970	1.515	0
	1	3.031	2.101	1.144	0
$\left(\frac{EL^j_*}{EL^e_*}\right)$	0	1	1	0.501	0
	0.2	1	0.891	0.454	0
	1	1	0.693	0.377	0
s^j_*	0	0.4000	0.4000	0.3416	0.2828
	0.2	0.4000	0.3885	0.3424	0.2944
	1	0.4000	0.4000	0.4000	0.4000

VI. Extensions

A. *Heterogeneous Commuters*

The extension to treat commuter heterogeneity is conceptually (though not algebraically) straightforward; indeed, the extension has already been done for the optimal fine-toll and no-toll equilibria with inelastic demand by Gordon F. Newell (1987), Cohen (1987), and Arnott et al. (1989), who consider heterogeneity in α, β, γ, and t^*. The essential insight is that in equilibrium user types follow a specific departure order; for example, "assembly-line workers" (those with a high γ) depart so as to arrive early. Consequently, one cannot simply apply the analysis of the paper to an average user.

With heterogeneous commuters, the results are not as neat. The simple reduced-form diagram does not extend; the relationship between aggregates is no longer as simple; and the characteristics of the various toll equilibria, as well as the generalizability of the MHS self-financing results, may depend on whether the congestion toll is anonymous.

While user heterogeneity undermines the simplicity of the results presented in the paper, it reinforces the basic message. In order to model user heterogeneity correctly, it is necessary to model different groups' departure-time decisions and, hence, to adopt a structural approach.

B. *Other Congestible Facilities*

We cast our analysis in terms of a very specific congestible facility: a bottleneck on a point-input, point-output road in the morning rush hour. An obvious question is to what extent our results generalize to other congestible facilities. The short answer is that there are advantages to modeling all congestible facilities structurally and explicitly treating users' decisions and the congestion technology, but the details of the analysis will differ from one congestible facility to the next. We have seen, for a specific case, that the standard model is poorly specified. Structural modeling imposes the discipline to ensure proper specification. The standard procedure of dividing the period of use into intervals and solving separately for equilibrium in each interval may, however, provide a good approximation for some congestible facilities.

The nature of users' decisions will differ according to the congestible facility. For all congestible facilities, time of use is an important margin of user choice. Other margins may be important too: with flow congestion in transportation, the cost functions depend on the vehicle headway drivers choose (Julio J. Rotemberg, 1985); a realistic model of parking would treat the decision concerning length of time parked (Amihai Glazer and Esko Niskanen, 1990); in telephone traffic, the user decides on call duration; in the use of public facilities, such as a swimming pool or zoo, visit length is important; in an art gallery, the extent of congestion depends on the distance viewers stand from pictures; and in wilderness areas, the cost functions depend on hikers'

choice of path and speed of travel (Mordechai Shechter and Robert C. Lucas, 1978).

The nature of the congestion technology also depends on the facility, and for most facilities there are several congestion-prone elements of capacity. For example, in going to a baseball game, a spectator will cruise for a parking spot, walk from the parking spot to the entrance, queue for a ticket, experience crowding while watching the game, and then after the game join a car queue to exit the parking lot. The extent to which congestion is dynamic (i.e., to which cumulative usage effects are important) also varies across facility types.

It remains to be seen how the details of the analysis differ according to the type of congestible facility. An analysis similar to that of this paper has been undertaken in de Palma and Arnott (1990), but for telephone congestion on a single line. In that study, modeling of telephone congestion is the same as that for the road bottleneck, except that (i) the queue discipline is such that customers in the queue are served on a random basis, while on the road they are served on a first-come first-served basis, and (ii) telephone capacity is a stock (the number of users that can be talking on a line *at a point in time*) while road capacity is a flow. The reduced-form formulas obtained were very similar to those in the present study. This demonstrates that the approach taken in the paper can be adapted to other congestible facilities. However, the extension to other congestible facilities for which there are multiple elements of capacity or for which congestion takes a form other than queueing will not be trivial.

We now comment on the extent to which the results recorded in the propositions generalize. Proposition 1 recorded the congestion cost function in our model under various pricing regimes. The results are specific and do not generalize. Proposition 2 compared the various tolling equilibria. The result that the efficiency loss is larger the less sensitive the tolling regime is obviously general. The other major result concerned optimal capacity: with $\varepsilon < 1$ (> 1) optimal capacity was larger (smaller) the less sensitive the tolling regime. These results generalize as follows. Since the tolls are set optimally conditional on the tolling regime, with identical individuals, each user faces a price equal to marginal social cost, and the envelope theorem holds. Consequently, the marginal benefit from capacity expansion is $-\partial \mathrm{TC}^j/\partial s$, and as long as marginal cost cuts marginal benefit from below, optimal capacity is larger under regime j' than under regime j if $-\partial \mathrm{TC}^{j'}/\partial s > -\partial \mathrm{TC}^j/\partial s$ for all s. With $\varepsilon < 1$, $-\partial \mathrm{TC}^j/\partial s$ should normally be larger the less sensitive the tolling regime, but the precise conditions will depend on the facility. Proposition 3 related the self-financing results. Equation (52) generalizes, provided each user faces a trip price equal to marginal social cost.

Our analysis suggests that previous studies have significantly underestimated the gains from the sophisticated tolling of urban roads. Those studies were based on models that ignored both cross-price effects and the dynamic nature of congestion and, hence, failed to capture the change in the distribution of departure times induced by tolls. Studies of other congestible facilities have generally treated cross-price effects but not the dynamic nature of congestion. Thus, it is unclear whether structural models of those facilities would uncover significant previously unperceived gains from more sophisticated pricing policies.[12]

[12]Compared to roads, the pricing of most other congestible facilities (e.g., telephone, computer, electrical, and water networks) is relatively sophisticated. Time-dependent pricing is the rule for long-distance telephone traffic and is not uncommon for electrical power. Priority pricing (whereby jobs with a higher priority move up the queue more quickly) is the rule on mainframe computers. Usage-dependent pricing is employed in some contexts; the University of Cambridge employs it for computer usage, and Hydro Québec employs it in some of its electrical pricing. Other forms of peak-load pricing include priority-service pricing (Robert Wilson, 1989) and demand-layer pricing (Robert S. Main, 1973). In all contexts, however, more sophisticated pricing is possible. For example, long-distance telephone rates could presumably be made very sensitive to the level of usage: a customer would dial a long-distance number, the efficient tariff based on current system usage would be displayed, and she would then decide whether to proceed with the call.

VII. Conclusion

In the Introduction we argued that the standard model of peak-period traffic congestion is poorly specified because it fails to model commuters' departure-time decisions and the congestion technology. To our knowledge, the first paper in the literature that addressed these deficiencies was Vickrey (1969). He constructed a model of the morning commute on a road with a single bottleneck in which a fixed number of identical individuals wish to arrive downtown at the same time and choose when to depart, trading off schedule delay against travel time. The queue length adjusts until, in equilibrium, trip price is uniform over the peak period. The solution to the model provides a function relating trip price and marginal social cost to capacity, the number of users over the entire rush hour, and the form of pricing. This yields the correctly specified reduced-form supply and cost functions. The trip demand functions, too, should be specified over the entire rush hour. Thus, with identical commuters, the version of the standard model which treats the period of use as a single interval is sound. However, the procedure employed in much of the literature, of dividing up the period of use into intervals and solving separately for equilibrium in each interval, is logically flawed.

In this paper, we examined some of the economic implications of the Vickrey model and also extended it to treat elastic demand and optimal capacity under a variety of pricing regimes. Two particular findings merit emphasis. First, with identical individuals, the MHS results on the degree of self-financing of congestible facilities with optimal capacity and optimal tolls were shown to apply regardless of the tolling regime. Hence, if a road of optimal capacity should be self-financing with an optimal time-varying toll, it should also be self-financing when only a gasoline tax (set at the optimal level) can be employed. Second, we computed the gains from efficient pricing to be considerably greater than those given in the empirical literature on urban auto congestion. The reason is that previous empirical estimates are based on a model that ignores the efficiency gain that results from a toll's ability to redistribute travelers over the rush hour. This suggests that more serious consideration should be given to sophisticated tolling schemes on urban roads.

While the analysis focused on rush-hour auto congestion, we discussed the generalization of our approach to other congestible facilities. We argued that there are always advantages to providing a *structural* model of a congestible facility, which explicitly treats users' decisions and the facility's congestion technology.

The appropriate directions for future research are evident. Vickrey's (1969) paper, supplemented by this one, provides the *method* for determining structural models of facilities subject to peak-load congestion. However, the modeling of the congestion technology is primitive. What we need now are more realistic models of congestion technologies and of consumers' behavioral decisions, along with empirical estimation of them. In the context of rush-hour traffic congestion, for example, models should be developed which derive hypercongestion (traffic-jam situations) from driving behavior, solve for equilibrium on a congested network, and account for heterogeneity among users in addition to accident and road damage costs. As well, it is clearly desirable to apply Vickrey's (1969) approach to other types of congestible facilities, notably computers, public utilities, airports, telecommunications, and recreational facilities.

REFERENCES

Aigner, Dennis, J. and Hirschberg, Joseph G., "Commercial/Industrial Customer Response to Time-of-Use Electricity Prices: Some Experimental Results," *Rand Journal of Economics*, Autumn 1985, *16*, 341–55.

Arnott, Richard, de Palma, André and Lindsey, Robin, "Bottleneck Congestion with Elastic Demand," Discussion Paper 690, Institute for Economic Research, Queen's University, 1987.

_____, _____ **and** _____, "Schedule Delay and Departure Time Decisions with Heterogeneous Commuters," *Transportation Research Record*, 1989, *1197*, 56–67.

_____, _____ **and** _____, "Economics of a Bottleneck," *Journal of Urban Economics*, January 1990, *27*, 111–30.

Arnott, Richard and Kraus, Marvin, "The Ramsey Problem for Congestible Facilities," National Bureau of Economic Research (Cambridge, MA), Technical Working Paper No. 84, 1990.

Ben-Akiva, Moshe, de Palma, André and Kanaroglou, Pavlos, "Dynamic Model of Peak Period Traffic Congestion with Elastic Arrival Rates," *Transportation Science*, May 1986, *20*, 164–81.

Boiteux, Marcel, "La Tarification des Demandes en Pointe: Application de la Théorie de la Vente au Coût Marginal," *Revue Générale de l'Électricité*, August 1949, *58*, 321–40; reprinted in English translation, "Peak-Load Pricing," *Journal of Business*, April 1960, *33*, 157–79.

Borins, Sanford F., "The Political Economy of Road Pricing: The Case of Hong Kong," in *Proceedings of the World Conference on Transport Research, Vancouver, BC*, Vol. 2, Vancouver: Center for Transportation Studies, University of British Columbia, 1986, pp. 1367–78.

Braeutigam, Ronald R., "Optimal Policies for Natural Monopolies," in Richard L. Schmalensee and Robert D. Willig, eds., *Handbook of Industrial Organization*, Amsterdam: North-Holland, 1989, pp. 1289–1346.

Braid, Ralph M., "Uniform versus Peak-Load Pricing of a Bottleneck with Elastic Demand," *Journal of Urban Economics*, November 1989, *26*, 320–7.

Cohen, Yuval, "Commuter Welfare under Peak-Period Congestion Tolls: Who Gains and Who Loses?" *International Journal of Transport Economics*, October 1987, *14*, 238–66.

Crew, Michael A. and Kleindorfer, Paul R., *The Economics of Public Utility Regulation*, Cambridge, MA: MIT Press, 1986.

de Palma, André and Arnott, Richard, "Usage-Dependent Peak-Load Pricing," *Economics Letters*, 1986, *20* (2), 101–5.

_____ **and** _____, "The Temporal Use of a Telephone Line," *Information Economics and Policy*, 1990, *4* (2), 155–74.

d'Ouville, Edmond L. and McDonald, John F., "Optimal Road Capacity with a Suboptimal Congestion Toll," *Journal of Urban Economics*, July 1990, *28*, 34–49.

Downs, Anthony, "The Law of Peak-Hour Expressway Congestion," *Traffic Quarterly*, July 1962, *16*, 393–409.

Dupuit, Jules, "De l'Influence des Péages sur l'Utilité des Voies de Communiction," *Annales des Ponts et Chaussées: Mémoires 1*, 1849, *817*, 170–248; translated by Elizabeth Henderson, "On Tolls and Transport Charges," *International Economic Papers*, 1962, *11*, 7–31.

Glazer, Amihai and Niskanen, Esko, "Parking Fees and Congestion," *Regional Science and Urban Economics*, March 1992, *22*, 123–32.

Hendrickson, Chris and Kocur, George, "Schedule Delay and Departure Time Decisions in a Deterministic Model," *Transportation Science*, February 1981, *15*, 62–77.

Keeler, Theodore E. and Small, Kenneth A., "Optimal Peak-Load Pricing, Investment, and Service Level on Urban Expressways," *Journal of Political Economy*, February 1977, *85*, 1–25.

Knight, Frank, "Some Fallacies in the Interpretation of Social Costs," *Quarterly Journal of Economics*, August 1924, *38*, 582–606.

Kraus, Marvin, "The Welfare Gains from Pricing Road Congestion Using Automatic Vehicle Identification and On-Vehicle Meters," *Journal of Urban Economics*, May 1989, *25*, 261–81.

_____, **Mohring, Herbert and Pinfold, Thomas,** "The Welfare Costs of Nonoptimum Pricing and Investment Policies for Freeway Transportation," *American Economic Review*, September 1976, *66*, 532–47.

Main, Robert S., "Periodic vs. Demand-Layer Pricing for Utility Loads," Ph.D. dissertation, University of California at Los Angeles, 1973.

McFadden, Daniel, "The Measurement of Urban Travel Demand," *Journal of Public Economics*, November 1974, *3*, 303–28.

Mohring, Herbert and Harwitz, Mitchell, *High-*

way Benefits, Evanston, IL: Northwestern University Press, 1962.

Newell, Gordon F., "The Morning Commute for Non-identical Travellers," *Transportation Science*, May 1987, *21*, 74–88.

Pigou, Arthur C., *The Economics of Welfare*, London: Macmillan, 1920.

Pucher, John and Rothenberg, Jerome, "Pricing in Urban Transportation: A Survey of Empirical Evidence on the Elasticity of Travel Demand," mimeo, Department of Economics, Massachusetts Institute of Technology, 1976.

Rotemberg, Julio J., "The Efficiency of Equilibrium Traffic Flows," *Journal of Public Economics*, March 1985, *26*, 191–206.

Schechter, Mordechai and Lucas, Robert C., *Simulation of Recreational Use for Park and Wilderness Management*, Baltimore: Johns Hopkins University Press, 1978.

Small, Kenneth A., "The Scheduling of Consumer Activities: Work Trips," *American Economic Review*, June 1982, *72*, 467–79.

_____, "The Incidence of Congestion Tolls on Urban Highways," *Journal of Urban Economics*, January 1983, *13*, 90–111.

_____, *Urban Transportation Economics*, Chur, Switzerland: Harwood, 1991.

Steiner, Peter O., "Peak Loads and Efficient Pricing," *Quarterly Journal of Economics*, November 1957, *71*, 585–610.

Strotz, Robert H., "Urban Transportation Parables," in Julius Margolis, ed., *The Public Economy of Urban Communities*, Washington, DC: Resources for the Future, 1965, pp. 127–69.

Vickrey, William, "Pricing in Urban and Suburban Transport," *American Economic Review*, May 1963 (*Papers and Proceedings*), *53*, 452–65.

_____, "Congestion Theory and Transport Investment," *American Economic Review*, May 1969 (*Papers and Proceedings*), *59*, 251–61.

_____, "Responsive Pricing of Public Utility Services," *Bell Journal of Economics*, Spring 1971, *2*, 337–46.

Walters, Alan A., "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, October 1961, *29*, 676–99.

Wheaton, William C., "Price-Induced Distortions in Urban Highway Investment," *Bell Journal of Economics*, Autumn 1978, *9*, 622–32.

Williamson, Oliver E., "Peak Load Pricing and Optimal Capacity under Indivisibility Constraints," *American Economic Review*, September 1966, *56*, 810–27.

Wilson, John D., "Optimal Road Capacity in the Presence of Unpriced Congestion," *Journal of Urban Economics*, May 1983, *13*, 337–57.

Wilson, Robert, "Efficient and Competitive Rationing," *Econometrica*, January 1989, *57*, 1–40.

Bureau of Labor Statistics, *Employment and Earnings*, U.S. Department of Labor, Washington, DC: U.S. Government Printing Office, 1991.

Entry, Dumping, and Shakeout

By RICHARD H. CLARIDA*

This paper studies the process of entry, dumping, and shakeout and the properties of the free-trade equilibrium that results in a world in which countries differ in their stock of technological knowledge, but not necessarily in their stock of technological capability. The model can account for the dumping of products by newly industrializing countries and technological newcomers at a price that fails to cover average cost in the face of a surge in world demand for the dumped product, a phenomenon that has not been uncommon in recent years.

> "[Between 1981 and 1986] the number of [dumping] petitions resulting in interventions... increased 258% [relative to the 1975–1980 period]. It is difficult to believe... that dumping has risen to more than two and one-half times its previous level especially since the strong dollar during the latter period made selling in the United States easier for foreigners rather than harder." [*Economic Report of the President*, 1988 p. 159]

Over the past decade, there has been an explosion in the number of dumping cases brought by U.S. producers against foreign exporters. The U.S. trade law defines dumping to occur when imports are sold in the United States at a price that is less than the home-country price (the traditional definition of dumping) or when the import price is less than the "constructed value" or average cost of production of the imported good (since a 1974 change in the law). In fact, in recent years, formal dumping complaints have increasingly often been based upon allegations of "pricing below cost" and not on allegations of price discrimination in favor of U.S. consumers (Howard Gruenspecht, 1988).[1]

In policy discussions, dumping is often attributed to a situation in which too many producers are chasing too few consumers.[2] In a cyclical downturn, considerations such as exit costs and implicit labor contracts may very well make it optimal and *ex ante* rational for existing firms to continue to produce at a price that falls short of average total cost.[3] Although the severity of the 1981–1982 worldwide recession no doubt can account for some of the surge in dumping findings during this decade, the fact that

*Columbia University and the National Bureau of Economic Research. This is a substantially revised version of an earlier paper entitled, "Destructive Competition, Dumping, and Learning by Doing." I thank Ron Findlay, Jagdish Bhagwati, Ricardo Cabellero, Maurice Obstfeld, Avinash Dixit, Ariel Pakes, two referees, and seminar participants at Columbia, Yale, Pennsylvania, Chicago, the NBER Summer Institute, UCLA, and New York University for their comments and suggestions, many of which have found their way into this paper. I am responsible for all remaining errors.

[1]This is not to say that international price discrimination on the part of exporters with market power, Jacob Viner's (1923) definition of dumping, is uncommon or unworthy of study. Rather, formal dumping complaints, at least in the United States, are increasingly often based upon allegations of "pricing below cost" regardless of transactions prices in home markets (Gruenspecht, 1988). Furthermore, "in testing for the presence of dumping, the foreign transactions price cannot be used for comparison with the American sales price if the former is less than the cost of production...[which] must include general expenses equal to at least 10 percent of manufacturing costs plus a profit margin of at least 8 percent" (Wilfred Ethier, 1982).

[2]For a superb account of the political economy of dumping, see Jagdish Bhagwati (1986).

[3]This was first shown rigorously by Ethier (1982), although the idea of cyclical dumping dates back to at least Viner (1923). In the conclusion, I briefly discuss the recent literature on strategic-duopoly models of dumping.

the number of preliminary findings of dumping by the United States Department of Commerce approached or exceeded all-time highs in the later years of the 1982–1990 economic expansion indicates that cyclical factors, as they are normally understood, cannot provide a complete explanation of dumping during the 1980's (*Annual Report on the Operation of the Trade Agreement Program*, various issues).

In this paper, I bring together the Ricardian model of international trade with a simple version of Boyan Jovanovic's (1982) model of entry, selection, and the evolution of industry to study one mechanism by which dumping can occur. The model developed in this paper can account for dumping cases in which foreign producers sell at a world price that fails to cover average cost when there is a surge in demand for the dumped product and a rise in total production of the dumped product in the dumping country. As I document in Section V, the conjunction of dumping at less than constructed value with a rise in the production and demand for the dumped product has not been an uncommon occurrence in recent years. It is, however, difficult to reconcile with traditional explanations of cyclical dumping, explanations that in fact account for dumping at less than average cost in the face of a fall in production and demand for the dumped product.

The novel feature of the two-sector, two-country Ricardian model developed in this paper is that, while entry is endogenous, firms can only learn about their production function for producing one of the goods by actually producing that good (as in Jovanovic [1982]). Furthermore, production functions, as indexed by a Ricardian labor productivity parameter, differ across firms. However, because producers are *ex ante* identical, all producers share a common entry decision rule based upon their knowledge of the population distribution of productivities across firms, the terms of trade, and their rational expectation of the future terms of trade. In turn, the equilibrium terms of trade depend upon the entry decisions made by firms, so that in a rational-expectations equilibrium, entry decisions and the terms of trade are mutually consistent functions of the parameters that describe technology and demand.

Endogenizing entry substantially enriches the conditions under which dumping can occur under perfect competition and rational expectations. Intuitively, dumping can arise because the present value of entry includes the value of information about future profitability. Such knowledge is valuable because the most productive producers earn positive current and future rents in equilibrium. Competition among *ex ante* identical producers for these present and future rents can drive down the price below the opportunity cost of the *ex post* efficient and inefficient firms who enter in search of the rents that accrue to the most efficient among them. I model a world of two countries that differ only in their initial stock of knowledge about the technology for producing one of the goods: the countries do not differ in the availability or distribution of this technology among firms, nor do they differ in size or tastes. Nonetheless, I show that if world demand is sufficiently high, dumping can occur in equilibrium. Interestingly, I also show that if world demand is sufficiently low, dumping cannot occur.

High demand induces entry of firms seeking to capture the rents accruing to the most efficient (*ex post*) among them. Entry among *ex ante* identical firms pushes down the world price below the opportunity cost of new entrants, and these firms will in fact initially export in equilibrium. *Ex post* high-cost producers recognize their losses and exit. The rents that accrue to the remaining producers in the dumping country exactly equal the net losses incurred as a result of dumping. Despite the fact that high world demand induces entry to the point that price is driven below private opportunity cost, which is the output per entrant forgone by (or average cost of entry to) the economy, I show that welfare in both the dumping (exporting) country and the importing country improve in the free-trade dumping equilibrium relative to autarky.

The plan of the paper is as follows. In Section I, I set up the model, and Section II derives the optimal entry decision rule. In

Section III, I specialize the model and obtain a closed-form expression for the autarky *ex ante* entry price in terms of the parameters of demand and technology. In Section IV, I provide explicit sufficient conditions for the existence of a dumping equilibrium in a two-country world in which the countries differ only in their initial stock of technological knowledge, not their technological capability, their size, or their preferences. In this equilibrium, positive trade flows arise from the (intertemporal) profit-maximizing efforts of firms in one of the two countries to accept a price below opportunity cost in exchange for the option of earning positive rents that accrue to the *ex post* most efficient among them. I also show that, while dumping coincides with excessive entry (excessive in that new entrants on average suffer losses), the entry that does occur maximizes social surplus subject to the constraint that productivity must be learned via production. Section V explores some of the implications of the analysis and surveys some recent dumping cases in which technology and a surge in world demand for the dumped product appeared to play a prominent role. Section VI provides concluding remarks and relates the results to those obtained elsewhere in the literature.

I. Setup of the Model

I consider a two-good, two-period, two-country Ricardo-Jovanovic world. Each country is populated by $N + M$ entrepreneurs. Each entrepreneur supplies inelastically one unit of labor, the sole factor of production. The two goods, z and y, are produced by entrepreneurs with constant-returns-to-scale production functions. I shall refer to entrepreneurs and firms interchangeably. In both countries, all $N + M$ entrepreneurs are endowed with one unit of labor and a technology that can be used to produce w units of y. N entrepreneurs in each country are endowed not only with one unit of labor and a technology for producing y, but also with a technology that can be used to produce z. Either w units of y or ξ^i units of z, but not both, can be produced in a given period by each potential entrant, $i = 1, \ldots, N$, to the z sector;[4] ξ^i differs among firms and is distributed according $G(\xi)$, which is common across both countries. Each entrepreneur knows this distribution but cannot learn the value of his particular ξ without first producing z.

I rule out borrowing and lending so that consumption must be financed by the value of current output. Let p (p') denote the present (future) price of z in terms of y; p and p' are endogenous in equilibrium but are taken as given by firms. Finally, I assume that preferences are time-separable and Cobb-Douglas. Let s denote the share of spending that falls on z by the N entrepreneurs that can produce z or y in each country, and let α denote the share of spending that falls on z by the M entrepreneurs that can only produce y in each country.[5]

II. The Entry Decision Rule

To derive the optimal entry decision rule followed by the N potential entrants to the z sector, begin by observing that the static utility gained by producing z in the present is $s^s(1-s)^{1-s}(p^{1-s})\xi^i$ while the static utility obtained by entering sector y is $s^s(1-s)^{1-s}(p^{-s})w$. Entry into sector z will maximize expected current utility whenever

$$(1) \qquad (p^{1-s})E\xi^i > (p^{-s})w.$$

Otherwise, static utility is maximized by entering sector y. Of course, to the extent that the entry decision in the present impacts upon expected utility in the future, (1) is not the optimal first-period entry decision rule. However, by substituting p' for p, (1) *does* describe the optimal second-period decision rule for firms that did not enter sector z in

[4] Production of z or y by agent i in the present in no way changes the parameters of the of technologies available for producing either z or y in the future (i.e., $\xi^i = \xi^{i\prime}$ and $w = w'$ in the future.

[5] Letting primes denote future values, entrepreneurs who can produce x or y maximize $z^s y^{1-s} + z'^s y'^{1-s}$ while those who produce y maximize $z^\alpha y^{1-\alpha} + z'^\alpha y'^{1-\alpha}$.

period 1. Rearranging, one obtains the following proposition.

PROPOSITION 1: *For all firms that produce y in period 1, entry to sector z in period 2 will occur if*

$$p' > w/E\xi^i \tag{2}$$

where $w/E\xi^i$ is the ex ante static break-even price of z, the price at which entrepreneurs achieve the same expected utility in either sector z or sector y.

What is the expected second-period utility of an entrepreneur who entered sector z in period 1? Such an entrepreneur learns his productivity in producing z. With this knowledge he can attain utility $s^s(1-s)^{1-s}(p'^{-s})w$ if he switches to sector y, and he can attain utility $s^s(1-s)^{1-s}\times(p'^{1-s})\xi^i$ if he stays in sector z. Note that, as of the beginning of period 2, ξ^i is not random, so that

$$s^s(1-s)^{1-s}E\max\left[(p'^{1-s})\xi^i,(p'^{-s})w\right] \tag{3}$$

is the expectation in the present of the utility in the future of an entrepreneur who enters the z sector in the present.

By contrast, it is easily seen from (1) and Proposition 1 that the expectation in the present of the utility in the future for an entrepreneur who enters the y sector in the present is given by

$$s^s(1-s)^{1-s}\times\max\left[(p'^{1-s})E\xi^i,(p'^{-s})w\right]. \tag{4}$$

The difference between (3) and (4) reflects the net expected value of information about the ability to produce z. Is this value positive? Since $\max[\cdot]$ is a convex function, by Jensen's inequality, the answer is given as the following proposition.

PROPOSITION 2:

$$E\max\left[(p'^{1-s})\xi^i,(p'^{-s})w\right]\geq\max\left[(p'^{1-s})E\xi^i,(p'^{-s})w\right]. \tag{5}$$

The nonnegative value of information may be understood as follows. When $p' > w/E\xi$ in period 2, knowledge of ξ^i obtained as a result of early entry allows *ex post* low-productivity firms to avoid remaining in the z sector even though all first-period y-producers—including those of low productivity—will by (2) maximize expected utility by entering the z sector at this price. When $p' < w/E\xi$ in period 2, knowledge of ξ^i obtained as a result of early entry signals high-productivity firms to remain in the z sector even though no period-1 y-producer will enter the z sector at this price. Thus, early entry has value, an option value if you will, if it provides information that permits high-productivity (low-productivity) firms to take advantage of (avoid) producing z when the price is low (high) relative to the opportunity cost of resources.

Pulling together (1), (4), and (5), I am now in a position to obtain a general expression for the entry price at which firms are *ex ante* indifferent between entering sector z and entering sector y in period 1. Letting b denote the static *ex ante* break-even price,

$$b \equiv w/E\xi^i \tag{6}$$

and letting e denote the (forward-looking) *ex ante* entry price, one has the following proposition.

PROPOSITION 3: *The rational, ex ante entry price at which firms are indifferent between entering the z sector and entering the y sector is given by*

$$e = b-(1/E\xi^i)E\max\left[p'\xi^i,w\right]+(1/E\xi^i)\max\left[p'E\xi^i,w\right]. \tag{7}$$

Furthermore, $e \leq b$, with strict inequality whenever the net value of information is positive.

For future reference, note that (7) can be rewritten as

$$e = b-(1/E\xi^i)E\max\left[p'\xi^i-w,0\right]+(1/E\xi^i)\max\left[p'E\xi^i-w,0\right]. \tag{8}$$

I conclude this section by observing that the *ex ante* entry price is related to the variance or, more generally, the dispersion of the probability distribution of productivities across agents. Indeed, a sufficiently large mean-preserving spread of G around $E\xi^i$ must drive the *ex ante* entry price below the static break-even price.[6] For completeness I state the following corollary.

COROLLARY 1: *Given p', b, and $E\xi^i$ there exists a mean-preserving spread of G such that*

$$E\max[p'\xi^i - w, 0] > \max[p'E\xi^i - w, 0] \tag{9}$$

(i.e., the value of learning about productivity is positive) and

$$e < b \tag{10}$$

(i.e., the ex ante entry price is less than the static break-even price).

III. Entry and Demand

Proposition 3 provides an expression for the *ex ante* entry price in terms of the opportunity cost of resources, w, the moments of the distribution of productivity across agents, $G(\xi)$, and the expected future price of z, p'. In any interesting model of dumping, p' must be endogenous. In the present model, entry decisions in period 1 depend upon the expected future price of z. However, the future price of z depends upon the future supply of z, which depends upon entry decisions in period 1. In a rational-expectations equilibrium, second-period prices and first-period entry must be mutually consistent.

Although my primary interest is in providing an account of dumping in international trade, I begin by solving the model for the autarky equilibrium in each country. Later, these results are used to derive the conditions under which dumping (the export of goods at a common world price below the average cost of production) can occur. Moreover, if the two countries are identical in all respects and none of the potential z-entrants in either country knows its firm-specific productivity in producing z, autarky will be the free-trade equilibrium.

To solve for the rational-expectations equilibrium, it is useful to make several further assumptions. I assume that $s = 0$; that is, the share of spending by potential or actual z-producers that falls on z is zero. This assumption insulates the position of the demand curve for z from the entry and exit decisions made by the producers of z. Entry and exit will have a decisive influence on price, but only because these decisions shift the supply curve along a given demand curve.[7] I shall also assume that there are three productivity parameters ξ^i and that these are distributed among the N potential z-entrants in each country according to the following distribution G:

$$\begin{aligned} &\xi = 1: && \text{shared by } (1-2\pi)N \text{ agents} \\ &\xi = 1+\varphi: && \text{shared by } \pi N \text{ agents} \\ &\xi = 1-\varphi: && \text{shared by } \pi N \text{ agents} \end{aligned} \tag{11}$$

with $0 < \varphi < 1$ and $0 < \pi < \frac{1}{2}$, so that

$$E\xi = 1 \qquad \operatorname{Var}\xi = 2\pi\varphi^2. \tag{12}$$

[6] PROOF: If $\max[p'E\xi^i - w, 0] > 0$, a positive probability of a sufficiently low realization of $\xi^i = \underline{\xi}$ such that $p'\underline{\xi} < w$ must make $E\max[p'\xi^i - w; 0]$ greater than $\max[p'E\xi^i - w, 0]$. If $\max[p'E\xi^i - w, 0] = 0$, a positive probability of a sufficiently high realization of ξ^i must make $E\max[p'\xi^i - w, 0]$ positive.

[7] Jovanovic (1982) also assumes that the position of the demand curve is invariant to entry and exit decisions. Jovanovic's model is much more general in that he allows for normally distributed shocks, Bayesian learning over time about the production function, and an infinite planning horizon. What I purchase with my simpler setup is the ability to characterize completely the two-country free-trade equilibrium in a Ricardian world in which countries differ in their initial stock of technological knowledge and to show how dumping—which is the export, not just the sale, of goods at a price below average cost—can occur.

The third assumption I shall make is that N and M are large enough so that the frequency distribution of productivities among firms that actually enter the z sector in period 1 is given by the population distribution.[8] It follows that the equilibrium p' is a function of the number of firms that enter in period 1, the population distribution of productivities, and the parameters of the demand curve for z, which for the Cobb-Douglas preferences is given by

$$(13) \qquad d_z(p) = M\alpha w / p.$$

It will be noted that, for the distribution in (11), the static break-even price is equal to the opportunity cost of resources w, $b = w$. These assumptions are sufficient to characterize completely both the autarky and free-trade equilibria. I begin with the following corollary.

COROLLARY 2: *Under the above three assumptions, the ex ante entry price is*

$$(14) \quad e = w - \pi(1-\varphi) \times \max[w/(1-\varphi) - p', 0]$$

so long as $p' \geq w$. *If* $p' \leq w$, *the ex ante entry price is given by*[9]

$$(15) \quad e = w - \pi(1+\varphi) \times \max[p' - w/(1+\varphi), 0].$$

It follows that the ex ante entry price is bounded by

$$w(1-\pi\varphi) \leq e \leq w.$$

Equation (14) says that early entry has value, so that e is driven below break-even level w, if it provides information that permits low-productivity firms to avoid producing z when the future price of z is high relative to the opportunity cost of resources. Equation (15) says that early entry has value, so that e is driven below break-even level w, if it provides information that permits high-productivity firms to take advantage of producing z when the future price of z is low relative to the opportunity cost of resources. In other words, early entry has value if it conveys knowledge that results in experienced firms making production decisions in the future that differ from the production decisions of new entrants in the future.

For ease of exposition, I shall confine my investigation of the phenomenon of entry at a price below average cost (and later the phenomenon of dumping—the export of goods at a price that falls short of average cost) to equilibria in which the future price of z, p', equals or exceeds the break-even level w. A sufficient condition for this to occur in the closed economy is simply the following:[10]

$$(16) \qquad (1-\pi)/\pi > 2\varphi.$$

Note that since $\pi < \frac{1}{2}$, the restriction that $\varphi < \frac{1}{2}$ will insure that $p' \geq w$.

Corollary 2 does not pin down the conditions under which the equilibrium price of z in period 1 is driven below the *ex ante* break-even price w. The following propositions do.

PROPOSITION 4: *Given* $\varphi < \frac{1}{2}$, w, α, π, *and* M, *a sufficient condition in the closed economy for the ex ante entry price to be*

[8]This assumption is in no way essential to the paper's main results, because all of these results go through in the special case in which demand for z at the *ex ante* entry price exceeds the maximum supply of z, N. Of course in this case, the distribution of productivities among entrants is equal to the population distribution.

[9]The proof of (14) and (15) just requires plugging the distribution (12) into equation (7) and canceling terms.

[10]Since $e \geq w(1-\pi\varphi)$, $\alpha M/(1-\pi\varphi)$ is an upper bound on first-period demand for z, as well as an upper bound on the number of first-period entrants. A fraction π of these firms are high-productivity and are willing to remain in the z sector if $w/(1+\varphi) \leq p'$. The output produced by these firms is then $\alpha M\pi(1+\varphi)/(1-\pi\varphi)$. If second-period demand at the break-even price $p' = w$ exceeds this supply, $p' \geq w$. This is the condition that $\alpha Mw/w > \alpha M\pi(1+\varphi)/(1-\pi\varphi)$ or $1 - \pi > 2\pi\varphi$. Since $\pi \leq \frac{1}{2}$, $p' \geq w$ if $\varphi < \frac{1}{2}$.

driven below the break-even price, $e < w$, is the following:[11]

$$(17)\quad \alpha M(1-\varphi)/[1-\pi(1-\varphi)] < N.$$

First-period entry has an option value if experienced firms can avoid losses in the future that inexperienced firms may incur. A fraction π of new entrants in the future will incur losses if $p' < w/(1-\varphi)$. Equation (17) is just the condition that potential future supply of z be large enough relative to demand at the price $w/(1-\varphi)$ so that the equilibrium future price falls short of this level, $p' < w/(1-\varphi)$. Thus, first-period entry has a value in that it permits *ex post* low-productivity firms to avoid producing z when the price of z is high relative to opportunity cost w.

I now provide sufficient conditions for the equilibrium period-1 price to be driven below opportunity cost. This means that a fraction $1-\pi > \frac{1}{2}$ of the firms that enter suffer losses in period 1 and that these losses exceed the rents earned by the *ex post* most efficient producers.

PROPOSITION 5: *There exists a great enough number of potential entrants, $\underline{N}$ such that, for $N > \underline{N}$, the equilibrium price of z in the closed economy is driven below the break-even level, $p < w$. $\underline{N}$ is given by*

$$(18)\quad \underline{N} \equiv \max[\alpha M, \alpha M(1-\varphi)/\{1-\pi(1-\varphi)\}].$$

Furthermore, there exists an $N_e > \underline{N}$ such that, for all $N > N_e$, the period-1 price is driven down to its minimum value $p = w(1-\pi\varphi)$. At this price, the average loss in terms of opportunity cost suffered by a new entrant to the z sector is given by

$$(19)\quad (p-w)/w = -\varphi\pi < 0.$$

Furthermore, N_e is given by

$$(20)\quad N_e \equiv \alpha M/[1-\pi(1-\varphi)].$$

(See the Appendix for a proof.)

[11] I will establish that if *ex ante* identical producers can sign an Arrow-Debreu insurance contract, that contract will specify a first-period entry price of $p = e$.

Even though Proposition 4 implies that the entry price e is driven below the break-even price w, it does not imply that the *equilibrium* price p is driven below the break-even price. For this to be the case, the output produced by all potential entrants N must exceed demand at the break-even price. Combining this condition with Proposition 4, I obtain Proposition 5. The max operator's two terms arise because there are two requirements for the equilibrium price to be driven below the break-even price. A necessary condition is that $e < w$, the second term in (18). It is also necessary that supply be large enough relative to demand so that $p < w$. This will be true when $N > \alpha Mw/w$, the first term in (18).

It is easy to see that $e < w$ does not imply $p < w$. For example, suppose that demand at the break-even price equals the output of all potential entrants $\alpha M = N$ so that $p = w$. In period 2, the low-productivity firms, πN in number, drop out, so that period-2 output falls to $N - \pi N(1-\varphi)$. The period-2 price will rise; but if $p' < w/(1-\varphi)$, which will be true if

$$N > \alpha M(1-\varphi)/[1-\pi(1-\varphi)]$$

then $e < w$ by Proposition 4. This happens for $\varphi > \pi/(1+\pi)$. Thus, both terms are needed in (18) to insure that the equilibrium price be driven below the break-even price.

Since all firms share the identical entry decision rule, the aggregate supply curve is perfectly elastic at the *ex ante* entry price e until $z = N$, at which point it is completely inelastic. Recalling that e is the price at which firms are indifferent between entering the z sector or the y sector in period 1, nothing guarantees that all firms that are indifferent between producing z or y will actually produce z. In particular if $d_z(e) < N$, all N producers cannot produce at e. If producers are rational, they take into account this possibility when forming their decision rule. However, if when $d_z(e) < N$ firms are chosen randomly from among the N potential entrants so that a fraction

$d_z(e)/N$ is selected, then the optimal entry decision rule is the same.[12]

Before showing how a free-trade equilibrium with dumping can arise, I first briefly explore the relationship among the *ex ante* entry price, the number of firms that enter the market for z, and the number of firms that would have entered the market for z (and the price they would have received) had each firm known *ex ante* its particular productivity in producing z. I define a *destructive* (*deficient*) *competition equilibrium* as an equilibrium in which, because firms are *ex ante* identical but *ex post* different, too many (too few) firms enter, driving the price below (keeping the price above) the level that would have prevailed had each firm known *ex ante* its particular productivity in producing z. In a destructive competition equilibrium, price is driven below average (opportunity) cost in period 1. I also show that, for certain parameter values, not only does a deficient competition equilibrium exist in this model, but perhaps surprisingly, price is driven below opportunity cost in a deficient competition equilibrium.

[12]However, the interpretation is slightly different (cf. Avinash Dixit and Carl Shapiro, 1986): e becomes the price at which firms are indifferent between (i) entering a lottery with probability $d_z(e)/N$ of winning, in which the winners produce z and the losers produce y, and (ii) not entering the lottery and just producing y. The proof of this result is just algebraic manipulation of (8). However, one insight is required. The probability of winning the lottery in the second period, θ', for firms that did not enter in the first period is endogenous and influences e according to

$$e = \beta - (1/E\xi^i)E\max[p'\xi^i - w, 0]$$

$$+ (1/E\xi^i)\theta' \max[p'E\xi^i - w, 0].$$

Recall from Proposition 1 that firms will be indifferent between newly entering z in period 2 and staying in y when $p'E\xi = w$. This can be shown to imply that, when $p'E\xi > w$, θ' must equal 1 and that, when $\theta' < 1$, $p'E\xi - w = 0$. Thus, without loss of generality, equation (8) defines the entry decision rule. This is not to say that expected utility is invariant to θ' or θ given p and p', but rather that utility is changed in such a way that entry decisions are unaffected.

PROPOSITION 6: *A destructive competition equilibrium exists if*

$$(21) \quad \alpha M/[1-\pi(1-\varphi)] \leq N \leq \alpha M/\pi(1+\varphi).$$

Furthermore, $p < w$ in a destructive competition equilibrium. A deficient competition equilibrium exists if

$$(22) \quad \alpha M/\pi \leq N.$$

Furthermore, $p < w$ in any such deficient competition equilibrium.[13]

How can aggregate losses occur when the price of z is higher and fewer firms enter than under *ex ante* knowledge of firm-specific productivities? In this model, price is driven below opportunity cost in the first period because *ex ante* identical firms are willing to accept expected losses in the present in exchange for the future profits that will accrue to the most productive among them. The fact that the number of firms that actually enter in period 1 (recall that all N are *ex ante* indifferent, as discussed in footnote 12) is less than would be the case under *ex ante* knowledge of firm-specific productivities is irrelevant.

IV. Dumping in a Free-Trade Equilibrium

With the results obtained in Section III, it is now straightforward to provide sufficient conditions for dumping to occur in a free-trade equilibrium. Dumping will be said to

[13]PROOF: The number of producers who, under full information, would produce at the *ex ante* entry price $e = w(1-\pi\varphi) < w$, is πN. Their total output would be $\pi N(1+\varphi)$. The price they would receive is $\alpha Mw/[\pi N(1+\varphi)]$. When $e = w(1-\pi\varphi)$, the second inequality in (21) implies that total output and entry are greater at e (and e itself is less than the price that would prevail under full information) than with *ex ante* knowledge about firm-specific productivities. Since $\pi(1+\varphi) < 1-\pi(1-\varphi)$ for $\pi < \frac{1}{2}$, the set of N that satisfies (21) is nonempty. By Proposition 5, $p < w$. When $e = w(1-\pi\varphi)$, (22) implies that total output and entry are less at e (and e itself is higher) than under *ex ante* knowledge of firm-specific productivities. By Proposition 5, $p < w$.

occur when z is exported at a common world price that falls below the opportunity cost of production in the exporting country. For positive trade flows—and thus dumping—to occur, it must be the case that the two countries not be identical. As discussed earlier, the results in Section II characterize the autarky equilibrium in each country, which is the free-trade equilibrium when the two countries are identical. By Propositions 4, 5, and 6, the price of z may be driven below opportunity cost in period 1. However, since the two countries are identical, no trade flows between them will arise in equilibrium, and no dumping can be said to occur.

I begin with the symmetric case in which countries 1 and 2 differ in only one respect: each potential z producer in country 1 knows its firm-specific productivity in producing z, while no potential z entrant in country 2 knows its firm-specific productivity. Otherwise, the distribution $G(\xi)$ of productivities across firms, the number of producers of y and potential producers of z (M and N, respectively), and their preferences are identical. Thus it can be said that countries 1 and 2 differ only in their initial stock of technological knowledge and not in their initial stock of technological capability. Interestingly, not only will such a difference in the initial stock of technological knowledge induce trade flows between the two otherwise identical countries, but under sufficient conditions to be established now, dumping will occur in the free-trade equilibrium. That is, in the first period country 2 will export z to country 1 at a common world price that is driven below the average cost of production in country 2. In the second period, following the excessive entry and resulting destructive competition that transpires in period 1,[14] a rationally expected shakeout occurs: inefficient country-2 firms, discovering their inability to cover opportunity cost, exit the z sector. This exit of country-2 producers drives up the period-2 price of z and acts to reduce or eliminate country 1's trade deficit in z.

[14] In what follows, I show formally that if dumping occurs in the symmetric case, the equilibrium must be a destructive-competition equilibrium.

PROPOSITION 7: *Given π, φ, M, and α, there exists a symmetric free-trade dumping equilibrium if*

$$(23)\quad \underline{N} < \max\left[2/\{1+\pi(1+\varphi)\},\ (1-\varphi)/\{1-\pi(1-\varphi)\}\right]\alpha M < N < \alpha M/\pi(1+\varphi).$$

In this dumping equilibrium, country 1 imports z from country 2 in period 1 at a price p that falls below average cost w.

(See the Appendix for a proof.)

This is a striking result. It says that for any values of the distribution parameters π and φ, the expenditure share $\alpha > 0$, the "wage" (opportunity cost) w, and the position of the world demand curve as indexed by M (the number of z-consumers in each country), there exists a number, N, of potential z-producers in each country such that country 1 is net importer of z at a price that is driven below the break-even point for country-2 exporters. This is despite the fact that the two countries differ only in their initial stock of technological knowledge, and not in their capability.

Three conditions must be met if dumping, the *export* of goods at a price below average cost, is to occur. First, there must be some value to learning about productivity, and thus firms must be willing to suffer static losses so that $e < w$. This is just the condition that $p' < w/(1-\varphi)$. In this symmetric world in which countries only differ in their stock of knowledge, not their stock of technology, this condition is the same as discussed under Proposition 4: (world) demand at the price $p' = w/(1-\varphi)$, $2\alpha M(1-\varphi)$, must fall short of the minimum (world) supply at that price, $2N - 2\pi N(1-\varphi)$. This is the second term in $\max[\cdot]$. It is graphed in Figure 1 and is denoted by the demand curve D3. All demand curves to the left of D3 insure $p' < w/(1-\varphi)$ so that country-2 firms are willing to enter the z sector in period 1 at a price $e < w$. The equilibrium period-1 price must in fact fall short of the break-even price w. What is the world supply at $p < w$? It is the supply of the efficient country-1 firms, $\pi N(1+\varphi)$, plus the supply

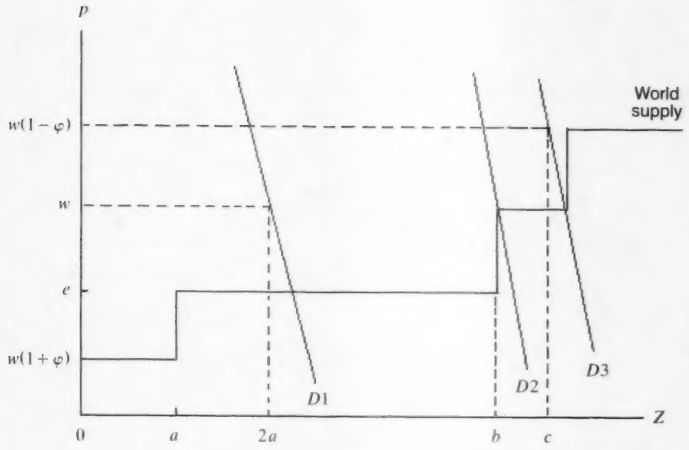


FIGURE 1. WORLD SUPPLY AND DEMAND

Notes: Supply a = $\pi N(1+\varphi)$; supply b = $N+\pi(1+\varphi)$; supply c = $2N-2\pi N(1-\varphi)$. New entrants in country 2 will be willing to sell z at a period-1 price $e<w$ if world demand lies to the left of D3. The equilibrium period-1 price p will fall below w if world demand lies to the left of D2. Country 2 will export z to country 1 if world demand lies to the right of D1 but to the left of D2. Thus, dumping will occur if world demand falls between D1 and D2.

of all country-2 entrants, N. Thus, $p<w$ if $2\alpha M<N+\pi N(1+\varphi)$. This is the firm term in max[·], which is graphed in Figure 1 and denoted by the demand curve D2. Thus, when the world demand curve lies to the left of D2 and D3, country-2 firms are not only willing to receive a period-1 price $p<w$, but do in fact receive it. As Figure 1 is drawn, π is sufficiently small so that $N+\pi N(1+\varphi)$, the period-1 world supply at $p\leq w$, is less than $2N-2\pi N(1-\varphi)$, the minimum period-2 supply at a price $w<p'<w/(1-\varphi)$.

Finally, there is a third requirement for dumping that is not present in the closed economy: the world demand at p must be more than twice as great as the supply of the efficient country-1 firms: $2\alpha Mw/p>2\pi N(1+\varphi)$. This condition is satisfied if the second inequality in (23) is met; it is graphed in Figure 1 as the demand curve D1. All demand curves to the right of D1 satisfy the requirement that the dumping country export the dumped good. Otherwise, even though firms in the "laggard" country sell below average cost, country 2 imports z from the country with the head start. That is, the low-cost producers in period 1 are the πN efficient firms in country 1 that are willing to produce at price $w/(1+\varphi)<e$. In this symmetric world, N cannot be too large, given demand, because large enough N implies that country 1 exports to country 2. However, this should not obscure the fact that, if one holds constant the number of country-1 firms and world demand, there exists a sufficiently large number of potential country-2 entrants such that dumping must occur.

COROLLARY 3: *The symmetric free-trade dumping equilibrium is a destructive-competition equilibrium.*[15] *If world demand is too low relative to potential supply, dumping can-*

[15]PROOF: If country 1 is to be a net importer of z, it must be the case that per-country demand at p exceeds country-1 supply, $\pi(1+\varphi)N$. Since country-1 supply is the supply of z in country 2 that would be forthcoming if country-2 producers had knowledge of firm-specific productivities, $\alpha Mw/p-\pi(1+\varphi)N$, "too many" country-2 firms enter the z sector in period 1.

not occur. Specifically, given N, M, π, and φ, there exists a value of α, $\underline{\alpha}$, such that if $0 < \alpha < \underline{\alpha}$, country 1 exports z to country 2 in period 1, and dumping does not occur; $\underline{\alpha}$ is given by

$$(24)\quad \underline{\alpha} = (1-\pi\varphi)\pi(1+\varphi)N/M < 1.$$

Alternatively, given N, π, φ, and α, there exists a value of M, $\underline{M}$, such that if $0 < M < \underline{M}$, country 1 exports z to country 2 in period 1, and dumping does not occur; $\underline{M}$ is given by the following:[16]

$$(25)\quad \underline{M} = (1-\pi\varphi)\pi(1+\varphi)N/\alpha.$$

COROLLARY 4: *The exit of inefficient country-2 producers drives up the period-2 price of z to $p' \geq w$ and acts to shrink the country-1 trade deficit in z. Furthermore, if*

$$(26)\quad N \geq \alpha M/[1-\pi(1-\varphi)]$$

such a shakeout in country 2 completely eliminates the country-1 trade deficit in z in period 2.[17]

While dumping of the sort investigated here cannot occur if demand is too low, it does not follow that, in the context of my model or a more general but similarly formulated model, that dumping must occur if demand is high. Given symmetry and N, there is a range in which demand may reside such that dumping occurs. If demand is too low, dumping cannot occur because demand can be met by "experienced" firms at a price that falls below the required entry price. This is Corollary 4. If demand is sufficiently high, given N, dumping can fail to occur for two reasons. First, p' can be driven above $w/(1-\varphi)$ so that there is no value to learning productivity this period: this is the second term in max[·]. Second, even if $p' < w/(1-\varphi)$ so that $e < w$, demand can be sufficiently strong so that the equilibrium $p \geq w$. This is the first term in max[·].

I now examine the welfare implications of dumping in a free-trade equilibrium. For concreteness, I focus on the case in which

$$(27)\quad \max[2/\{1+\pi(1+\varphi)\},\ 1/\{1-\pi(1-\varphi)\}]\alpha M < N < \alpha M/\pi(1+\varphi).$$

As can be seen by comparing (23) and (27), any N that satisfies (27) will satisfy Proposition 7. Furthermore, such an N will insure that $p' = w$ in autarky in both countries, that $p = w$ in autarky in country 1, and that $p = e = w(1-\pi\varphi)$ in country 2.[18] Under autarky, welfare (as measured by the sum of consumer and producer surplus in the market for z) is given by

$$(28)\quad 2\left(\int_w^\infty (\alpha Mw/P)\,dP + (w - w/(1+\varphi))\pi(1+\varphi)N\right)$$

in country 1, while welfare in country 2 is just

$$(29)\quad \int_e^\infty (\alpha Mw/P)\,dP + \int_w^\infty (\alpha Mw/P)\,dP.$$

Note that the most efficient producers in

[16] To see this, note that, since $p \geq w(1-\pi\varphi)$, $d_z(p) \leq 2\alpha M/(1-\pi\varphi)$. Thus, if it is the case that $2\alpha M/(1-\pi\varphi) \leq 2\pi(1+\varphi)N$, country 1 is a net exporter of z in period 1. Equations (24) and (25) follow immediately.

[17] This is not to say that country 2 cannot run a trade surplus in z in period 2. However, the conditions for this to occur will not be presented, although they are straightforward, if tedious, to derive. To see (26), note that if (23) is satisfied, $w \leq p' < w/(1-\varphi)$ so that αM is the maximum period-2 demand for z in country 1. $N[1-\pi(1-\varphi)]$ is the country-1 supply of z at any $w \leq p' < w/(1-\varphi)$. Thus, if (26) is satisfied, domestic production in country 1 can meet demand. Note finally that $\alpha M/[1-\pi(1-\varphi)] < \alpha M/\pi(1+\varphi)$, so that the set of N that satisfies (26) and (23) (and thus Proposition 7 and Corollary 4) is nonempty.

[18] PROOF: The country-2 results follow immediately from Proposition 5: $\{1/[1-\pi(1-\varphi)]\}\alpha M < N < \alpha M/\pi(1+\varphi)$ is just the condition that per-country demand exceeds the supply of the most efficient producers but falls short of the total supply of all but the least efficient producers.

country 1 earn positive rents in autarky, while the net producer surplus in country 2 is zero in autarky. This is a direct implication of the fact that the period-1 autarky price is driven down to the *ex ante* entry price in country 2, and that the equilibrium period-2 price of z is pushed up to the *ex ante* break-even level w.

Under the conditions of Proposition 7, a free-trade dumping equilibrium exists in which $e \leq p < w$ and, by (27), $p' = w$. The change in welfare in country 1 is given by

$$(30)\quad \int_p^w (\alpha Mw/P)\, dP - (w-p)\pi(1+\varphi)N > 0.$$

Despite the fact that dumping occurs in the free-trade equilibrium, country 1 unambiguously benefits from free trade. The loss in producer surplus resulting from the fall in the price of z represents a transfer from the most efficient producers to consumers. Since country-1 demand for z exceeds the output of the most efficient z-producers, the gain in consumer surplus must exceed the loss in producer surplus. Note that, relative to autarky, employment in the z sector is lower by $\alpha M - \pi(1+\varphi)N$, as less efficient z-producers, unable to cover their opportunity cost, shift to the y sector. Thus, both employment and profits in the country-1 z sector shrink in period 1 in a free-trade dumping equilibrium.

The change in welfare in country 2 is given by

$$(31)\quad \int_p^e (\alpha Mw/P)\, dP + (p-e)N \geq 0.$$

Welfare rises in country 2 if world demand is sufficiently strong to pull up $p > e$. Otherwise, country-2 welfare is unchanged in the free-trade dumping equilibrium. How can this be? The supply of z in country 2 is completely elastic at the *ex ante* entry price e. Country-2 producer's *ex ante* expected and *ex post* average rents are driven to zero at e. Although employment and production in the z sector increase in a free-trade dumping equilibrium (relative to the autarky equilibrium), only a fraction π of these exporters will obtain positive producer surplus *ex post*; the remaining fraction $1-\pi$ will suffer losses that exactly offset these gains. Obviously, country-2 consumers lose if world demand pushes p above e; but this loss must fall short of the gain to producers, because country-2 supply at $p > e$ is equal to N, while country-2 demand at e (and thus at p) is less than N. Thus, I have now established the following proposition.

PROPOSITION 8: *Consider the free-trade dumping equilibrium that satisfies (27). Then, relative to autarky, the following must be true:*

(a) The welfare of country-1 consumers improves; the welfare, rents, and period-1 output of country-1 z-producers contract; and the gain in country-1 consumer surplus exceeds the loss in producer surplus.

(b) The welfare of country-2 consumers does not improve; the welfare, rents, and period-1 output of country-2 z-producers do not contract; and the gain, if any, in country-2 producer surplus exceeds the loss, if any, of country-2 consumer surplus.

Note that while less efficient z-producers in country 1 are unable to cover their opportunity cost in period 1 of the dumping equilibrium, they only earn w in autarky. Thus the extent of injury in the z sector in country 1 is the loss of pure rents that accrue to the most efficient producers. It is easy to generalize this result. If N satisfies Proposition 7 such that the autarky price in country 1, p_A^1, satisfies $p_A^1 = p' > w$ and $p' > w > p_A^2 > e$ in country-2 autarky, then welfare in country 1 still improves in the free-trade dumping equilibrium:

$$\int_p^{p'} (\alpha Mw/P)\, dP - (p'-p)\pi(1+\varphi)N - (p'-w)(1-2\pi)N > 0$$

which can be shown to be positive if (23) is

satisfied. The change in country-2 welfare is given by

$$\int_{p}^{p_A} (\alpha Mw/P)\, dP + (p - p_A)N \geq 0$$

which is also positive. In this equilibrium the z-producers who cannot cover opportunity cost in the dumping equilibrium actually achieve lower utility than in autarky.

I will now discuss the robustness of these results. Consider the willingness of potential entrants to suffer expected static losses today in exchange for higher returns in the future. In my construct, information about productivity has value if and only if this information means that "established" firms make different production decisions in the future than do new entrants. Of course, the return to an established firm cannot be lower than the return to a new entrant, and it will in general be higher. Furthermore, there must be greater supply of z in the first period than in the future because some inefficient entrants exit; thus, given demand, prices must rise between period 1 and 2. Incorporating the value of information, new entrants in period 1 are willing to take static losses to obtain the experience that comes from producing. This insight seems quite general. However, even in my simple model, it is not always true that established firms earn higher expected returns than do entrants. If future price is sufficiently high, $p' \geq w/(1-\varphi)$, experienced and inexperienced firms choose the same sector, z, in the future. Since *ex ante* in period 0 experienced and inexperienced firms are the same, their expected period-2 returns are the same: there is no value to early entry. Thus, as I pointed out above, if demand is too strong dumping will not occur. However, there is a range in which the demand curve may reside such that demand is sufficiently high to induce entry and dumping, but not high enough to insure positive expected rents to all early entrants. These are exactly the parameter restrictions discussed at length above.

I conclude this section with an investigation of the welfare properties of the autarky equilibrium in a country in which each potential entrant is initially uncertain about its productivity in the z sector. At the beginning of period 1, all potential z-producers are *ex ante* identical. Suppose they sign an Arrow-Debreu insurance contract that guarantees that each z-producer will share equally in the total output of all z-producers in both periods. Let p^* denote the first-period price at which such a contract would require entry to the z sector. For any given vector of current and future prices (p,p') such that $w \leq p' \leq w/(1-\varphi)$ and $p \geq p^*$, the utility achieved by each potential z-entrant is given by

$$(32)\quad u(p,p') = \theta(p)p + [1-\theta(p)]w + \theta(p)p'[\pi(1+\varphi)+(1-2\pi)] + \theta(p)w\pi + [1-\theta(p)]p'$$

where $\theta(p) = \min[d_z(p)/N, 1]$. Note that a potential z-entrant will only agree to such an Arrow-Debreu contract if $u(p,p') \geq w + p'$, the expected value of producing y this period, entering the z sector next period, and keeping the returns which, in expectation, are equal to p'. It follows that the insurance contract will only require entry[19] to the z sector in period 1 if $p \geq p^*$ such that

$$(33)\quad \theta(p)p + \theta(p)p'[\pi(1+\varphi)+(1-2\pi)] + \theta(p)w\pi \geq \theta(p)w + \theta(p)p'$$

which is just the condition that p^* satisfies

$$(34)\quad p^* = w - \pi(1-\varphi) \times \max[w/(1-\varphi) - p', 0].$$

This is exactly the entry decision rule derived in Section III, equation (14). It follows that, for any given p', the amount of

[19] As discussed in Section III, when $d_z(p) < N$, it is assumed that all potential z entrants in fact enter a lottery in which winners produce z and losers produce y.

entry and, in a two-country world, dumping that occurs is exactly the same irrespective of the presence or absence of an Arrow-Debreu insurance market. The Arrow-Debreu contract will stipulate that all signees enter the z market when $p \geq p^* = e$. When $p = e$, each z-producer receives

$$e < \theta(e)e + [1 - \theta(e)]w < w$$

in period 1 (and of course each of the $[1 - \theta(e)]N$ signees who are y-producers also receive $\theta(e)e + [1 - \theta(e)]w$). With insurance, the loss to each z-producer is lower in the first period than is the expected loss to each z-producer in the absence of insurance. In period 2, each insured z-producer receives

$$(35)\quad \theta(p)p'[\pi(1+\varphi) + (1-2\pi)] + \theta(p)w\pi + [1 - \theta(p)]p'$$

which is less than the expected return, in the absence of insurance, of $p'[\pi(1+\varphi) + (1-2\pi)] + w\pi$ received by each experienced z-producer in period 2. The key point is that, with risk neutrality, the facility to pool risk afforded by an insurance contract does not alter conditions under which it is optimal to enter the z sector in the first period. Such a contract insures that each firm earns the same utility $u(p, p')$. In the absence of such a contract, the expected utility of each potential z entrant is $u(p, p')$.

Having demonstrated in Proposition 8 that, relative to autarky, welfare in the free-trade dumping equilibrium must improve, I now show that the autarky competitive equilibrium in fact maximizes the sum of consumer and producer surplus. Again, for concreteness I focus on the case in which rents are driven to zero in autarky equilibrium, which by Proposition 5 will be true if $N \geq \alpha M/[1 - \pi(1 - \varphi)]$. Since $p' = w$ in such an equilibrium, producer surplus at any first-period price p is given by

$$(36)\quad \text{ps}(p) = (p - w)d(p) + [w(1+\varphi) - w]\pi d(p).$$

Let $\tilde{p} < p$. Then the change in producer surplus may be expressed as

$$(37)\quad \text{ps}(\tilde{p}) - \text{ps}(p) = (\tilde{p} - p)d(p) + [\tilde{p} - w(1 - \pi\varphi)][d(\tilde{p}) - d(p)].$$

The change in consumer surplus is just

$$(38)\quad \text{cs}(\tilde{p}) - \text{cs}(p) = \int_{\tilde{p}}^{p} (\alpha Mw/P)\, dP > (p - \tilde{p})d(p).$$

Thus, for all $\tilde{p} < p$, social surplus at $\tilde{p}$ exceeds social surplus at p if

$$(39)\quad \tilde{p} > e = w(1 - \pi\varphi).$$

Now let $p = e$. If $\tilde{p} < e$, (37) implies that

$$(37')\quad \text{ps}(\tilde{p}) - \text{ps}(e) = (\tilde{p} - e)d(\tilde{p})$$

which means that producer surplus unambiguously contracts. Furthermore, since

$$(38')\quad \text{cs}(\tilde{p}) - \text{cs}(e) = \int_{\tilde{p}}^{e} (\alpha Mw/P)\, dP < (e - \tilde{p})d(\tilde{p})$$

social surplus at $\tilde{p} < e$ falls short of social surplus at e. Thus, I have just proved the following proposition:

PROPOSITION 9: *Consider an autarky equilibrium in which the expected rent to a potential entrant is driven to zero. This equilibrium, which will occur whenever* $N \geq \alpha M/[1 - \pi(1 - \varphi)]$, *maximizes the sum of consumer and producer surplus subject to the constraint that all potential entrants to the z sector are ex ante identical.*

V. Implications

In this section, I explore some key implications of the model of entry, dumping, and shakeout. In pursuit of this goal, it will be useful to generalize slightly the model of Section IV to allow for differences between

countries in w, productivity of labor in the y sector, and thus the opportunity cost of resources in the z sector. In particular, let $\omega < w$ denote the opportunity cost of resources in country 2 and suppose that

$$w/(1+\varphi) < \omega(1-\pi\varphi) < \omega < w.$$

I continue to assume that distribution of productivities across firms is identical in the two countries but that no potential entrant in country 2 knows its firm-specific productivity in the z sector until it has actually produced z. I begin by stating sufficient conditions for a free-trade dumping equilibrium to exist in such a world economy.

PROPOSITION 10: *Given π, φ, ω, w, M, and α, there exists a free-trade dumping equilibrium if (39) is satisfied and if:*

$$(40)\quad \{1/[1-\pi(1-\varphi)+\pi(1+\varphi)]\} \times \alpha M(\omega+w)/\omega < N < \alpha M(\omega+w)/\omega 2\pi(1+\varphi).$$

In this equilibrium, $p < \omega$, $p' = \omega$, country 1 imports z from country 2 in both periods, and dumping occurs in the first period.[20]

Because $\omega < w$ and the distribution $G(\zeta)$ of productivities across firms in the z sector is the same in both countries, country 2 has a comparative advantage in the z sector, and it will export z to country 1 in both periods. However, notwithstanding its comparative advantage in the z sector, country 2 can still dump in period 1 for the reason outlined in Sections II and III: firms are willing to enter the z sector at a price that fails on average to cover the opportunity costs of resources in period 1 if the most efficient among them earn rents in period 2.

Even with a comparative advantage in z, country 2 (the "low-wage" country) dumps in period 1 if the conditions of Proposition 10 are met. Again, three conditions must be met if dumping is to occur: there must be value to period-1 entry, $p' < \omega/(1-\varphi)$; demand must exceed twice the supply of the efficient country-1 producers (otherwise these producers export to country 2); and demand at the break-even price ω must not exceed potential world supply at that price. These are the parameter restrictions that are defined by Proposition 10. In fact, for concreteness, Proposition 10 sets out restrictions such that $p' = \omega$.

While country 1 continues to import z in period 2, there is shakeout of inefficient country-2 producers and a contraction in country-2 output in period 2 that pushes up the world price of z to $p' = \omega$.[21] I now show that, under a mild additional restriction on the parameter φ (the fraction by which the outputs of the most and least efficient firms differ from the output of the average firm), it is possible to obtain simple and intuitive expressions for the rate at which inefficient firms exit during the shakeout as well as the average loss incurred by new entrants in period 1.

PROPOSITION 11: *If $\varphi \leq \frac{1}{3}$, then in any free-trade dumping equilibrium that satisfies Proposition 10, the equilibrium price of z in period 1 is driven down to $p = \omega(1-\pi\varphi)$, the average loss incurred per new entrant in period 1 is equal to $-\pi\varphi\omega$, and the rate at*

[20] PROOF: $\alpha M(\omega+w)/\omega$ is the world demand for z at the country-2 break-even price ω. The minimum world supply at $p' = \omega$ is just $\pi(1+\varphi)N + (1-\pi(1-\varphi))N$ since, by (39), $\omega < w$. At any $p' < w$, $2\pi(1+\varphi)N$ is the maximum world supply of z. Equation (40) follows immediately as the condition for $p' = w$. For dumping to occur, $p < w$, which will be true if $\alpha M(\omega+w)/\omega < N + \pi(1+\varphi)$, as implied by the first inequality in (40). It must also be true that $\alpha Mw/\omega$ exceeds $\pi(1+\varphi)N$, as is implied by the second inequality in (40).

[21] However, because $w/(1+\varphi) < \omega(1-\pi\varphi) < \omega < w$, the rise in the world price of z does not call forth an expansion in the country-1 output of z, which in both periods is given by $\pi(1+\varphi)N$.

which inefficient country-2 firms exit during the shakeout is constant and equal to π.[22]

Proposition 11 highlights an interesting feature of the particular dumping equilibrium studied in this section; namely, given $1-2\pi$, the probability of being a firm of average productivity, the exit *rate* is independent of the average loss per entrant. The larger is φ, the fraction by which the output of the most (and least) efficient firms differs from the output of the average firm, the smaller will be p, and thus the greater will be the average loss per entrant. Since $p' = \omega$, all low-productivity entrants will exit as part of the shakeout. These low-productivity firms represent a fraction π of all period-1 entrants. The condition that $\varphi \leq \frac{1}{3}$ is just the condition that the output produced by each exiting firm, $1-\varphi$, is sufficiently great so that *only* low-productivity firms exit as part of the shakeout, and *ex post* all entrants of average and, of course, high productivity produce in both periods.

Thus, under the conditions of Proposition 11, some part of the world demand for z in period 2 will be met by new entry in country 2. This means that the change in country-2 output is the sum of two forces working in opposite directions: the shakeout of inefficient period-1 entrants and the entry of new z-producers in period 2. Of course, under the conditions of Proposition 11, the reduction in output resulting from the shakeout, in absolute value, must exceed contribution to output arising from the new entry. However, it follows that the contraction in output understates the amount of shakeout and dumping for two reasons. First, because of the entry of new z-producers in period 2, the contraction in output must be less than the output produced by inefficient firms that exit during the shakeout. Second, the number of firms that exit as part of the shakeout understates the number of firms that dump in period 1. This follows from the fact that p is driven down to $\omega(1-\pi\varphi)$, so that firms of average productivity, $(1-2\pi)N$ in number, also dump in period 1. Yet none of these firms exit in period 2. In other words, the fact that a firm does not exit as part of the shakeout does not imply that such a firm did not dump in the first period.

Proposition 11 does not rule out a positive correlation between the exit rate and the average loss per entrant. Indeed, given φ, both the exit rate and the average loss per entrant are increasing in π. Thus, the effect of a mean-preserving spread of $G(\zeta)$ is to raise both the exit rate and the average loss per entrant. Note finally that if π and φ are negatively correlated, so that sectors with few firms of average productivity are also sectors in which the outputs of the most and least efficient firms differ little from the output of the average firm, then the exit rate and the average loss per entrant can in fact be negatively correlated. In sum, without rather detailed knowledge of the distribution of productivities among potential entrants, a correlation of either sign, or of zero, between exit rates and average losses is consistent with some version of the model.

For a given world demand curve, I have shown that the contraction in output in the dumping country that follows the initial stage of entry and shakeout understates the amount of shakeout and dumping that actually occurs. I now generalize the model to allow for growth in world demand and show that it is possible for the initial stage of dumping and shakeout to be followed by an

[22] PROOF: Letting $D(p) \equiv \alpha M(\omega + w)/p$, then $\pi(1-\varphi)[D(\omega(1-\pi\varphi)) - \pi(1+\varphi)N]$ is the output produced by exiting firms at the period-1 price $p = \omega(1-\pi\varphi)$. The contraction in country-2 output between periods 1 and 2 is just $D(\omega(1-\pi\varphi)) - D(\omega)$. The output produced by exiting firms will exceed the contraction in country-2 output if

$$D(\omega) > \pi(1+\varphi)N(1-\varphi)(1-\pi\varphi)/(1-2\varphi).$$

From (40), this will be true if $(1-\varphi)(1-\pi\varphi)/(1-2\varphi) < 2$. Rearranging, this is just the condition that $3\varphi < 1 + \pi\varphi - \pi\varphi^2$, which for any $\pi > 0$, will be true if $\varphi \leq \frac{1}{3}$. If $D(\omega)/(1-\pi\varphi) < N(1+\pi(1+\varphi))$, $p = \omega(1-\pi\varphi)$. From (40), it is known that

$$D(\omega)/[1-\pi(1-\varphi)+\pi(1+\varphi)] < N.$$

It is easy to verify that, if $\varphi < \frac{1}{3}$, $(1-\pi\varphi)[1+\pi(1+\varphi)] > [1-\pi(1-\varphi)+\pi(1+\varphi)]$ and thus $p = \omega(1-\pi\varphi)$.

expansion in output in the dumping country. To see this, let $\alpha M'(\omega+w)/p'$ denote the world demand for z in period 2, and suppose for concreteness that $M'/M = (1+g)/(1-\pi\varphi)$. With this rate of world demand growth, the demand for z at the period-2 price of $p' = \omega$ exceeds the demand for z at the period-1 price $p = \omega(1-\pi\varphi)$. If N is such that

$$(41)\quad \{1/[1-\pi(1-\varphi)+\pi(1+\varphi)]\} \times \alpha M'(\omega+w)/\omega < N < \alpha M(\omega+w)/\omega 2\pi(1+\varphi)$$

is satisfied, then dumping occurs in period 1, and inefficient firms exit as part of the shakeout; but output expands after the shakeout in response to the fully anticipated rise in world demand.[23]

In addition to the above-mentioned implications, there are at least three additional insights conveyed by this approach that may be directly compared with the record so as to assess the relevance of the theory. First, following the lead of Michael Gort and Steven Klepper (1982), the model implies that dumping should be observed in younger industries in which new technology —new at least to the dumping country— plays an essential role. Second, the model implies that, even in countries in which an existing production base is in place, a surge in demand, to the extent that it attracts new entrants or new additions to capacity, may be associated with dumping. Third, the model implies that dumping should be observed in countries whose firms have not yet mastered an existing technology or acquired sufficient experience in producing the dumped good. I consider in turn the empirical support for each of these predictions.

Gort and Klepper (1982) develop an evolutionary theory of industry structure following the introduction of a product innovation.[24] According to their theory and the case studies of American industries they report, the markets "for most new products appear to pass through at least five distinguishable stages in the course of their evolution. New industries generally pass through a stage (IV) in which the number of producers declines significantly" (Gort and Klepper, 1982 p. 651). In Gort and Klepper's descriptive model, a new product is commercially introduced in stage I, followed by a stage II in which the entry rate and industry output accelerate rapidly. In stage III, the rate of *net* entry declines dramatically, ultimately to zero, as a result of "the accumulation of experience by existing firms...and a gradual reduction in the population of potential entrants that have not as yet entered the market" (Gort and Klepper, 1982 p. 633). In the aforementioned stage IV, the "less efficient firms who are unable to imitate the leaders...are forced out of the market.... Consequently, the exit rate rises sharply" (Gort and Klepper, 1982 p. 633). Finally, in stage V, the net entry rate returns to zero, where it remains until the eventual shrinkage of the market is brought on, either by product obsolescence or the introduction of a competing product.

Following the lead of Gort and Klepper (1982) and their case-study approach, I searched the record of dumping cases filed during the last decade to isolate those in which new technology and a substantial shift in world demand might be expected to play a prominent role. I confined my attention to cases in which the Commerce Department

[23] Note that the set of N that satisfies (41) is nonempty if

$$(1-\pi\varphi)/2\pi(1+\varphi) > \{1/[1-\pi(1-\varphi)+\pi(1+\varphi)]\}.$$

For any given π, there exists sufficiently small but positive φ and g such that (41) is satisfied.

[24] I thank a referee for bringing the Gort and Klepper (1982) paper to my attention. The literature on competitive, as opposed to strategic, models of selection in industries dates back at least to Armen Alchian (1950) and Richard R. Nelson and Sidney G. Winter (1982). Following on the seminal work of Jovanovic (1982) and Steven Lippman and R. Rumelt (1982), recent papers by Hugo Hopenhayn (1986), Ariel Pakes and Richard Ericson (1988), and Jovanovic and Saul Lach (1989) have begun to explore the implications of learning, entry, and exit on industry structure in a closed-economy context.

ruled that sales "at less than fair value" were in fact made. Within this subset of cases, I examined only those in which the Commerce Department employed the "constructed value" approach to determine fair value. Constructed value is the Department's estimate of the average cost of production. It is employed whenever there are insufficient sales (and many times no sales) in the domestic market that can be used to calculate the domestic sales price. Thus, a finding of dumping (sales at less than fair value) based upon constructed value is a finding that the dumped goods have been exported at a price falling short of the Commerce Department's estimate of average cost.

Among the cases I isolated in which technology and robust world demand played a prominant role, two stand out: the dumping of cellular telephones and cellular-telephone subassemblies by Japan in 1984; and the dumping of small-business telephone systems and small-business telephone subassemblies by South Korea in 1988. In 1985, in response to a complaint filed by the Motorola Corporation in November 1984, the Commerce Department ruled (*Federal Register*, Vol. 50 [1985] p. 24554) that Japanese manufacturers were dumping cellular telephones and cellular-telephone subassemblies in the United States. In its ruling, the Department noted that, because of insufficient sales in the domestic market by at least three firms, the finding of dumping was based on evidence that cellular telephones were being exported at a price that failed to cover average cost. In 1989, in response to a complaint filed by the AT&T and Comdial Corporations, the Commerce Department ruled (*Federal Register*, Vol. 54 [1989] p. 31980) that South Korean manufacturers were dumping small-business telephone systems and subassemblies in the United States. Again, the finding of dumping was based on evidence that these business telephone systems were being exported to the United States at a price that failed to cover average cost.

Steven Olley and Pakes (1988), in their recent study of the U.S. telecommunications industry, have documented that the demand for communication equipment, and especially imported equipment, surged after the breakup of AT&T in the early 1980's. In 1982, imports represented 4 percent of new supply in the U.S. market; by 1987 the share of imports had more than tripled to 14 percent. Olley and Pakes (1988) also document that this surge in demand for imports accompanied a substantial jump in the entry (and exit) rate of new plants, owned by domestic as well foreign producers, in the U.S. market. They argue that, "entry into the telecommunications equipment market was effectively prohibited...prior to the AT&T divestiture [since] Western Electric supplied 90 percent of AT&T equipment purchases" (Olley and Pakes, 1988 p. 4). Furthermore, it was not until 1978 that the Federal Communications Commission even permitted the connection of such private subscriber equipment as telephones and PBX systems to the AT&T network. While I was not able to obtain detailed data on entry and exit rates at the plant level for the Japanese cellular-telephone industry or the South Korean business-telephone-system sector, it is noteworthy that the constructed-value definition of dumping was employed by the Commerce Department precisely because of insufficient home-market sales of these products; price discrimination was neither alleged nor documented. Furthermore, because of the surge in demand for telecommunications equipment following the AT&T divestiture, these cases would not appear to represent episodes of cyclical dumping.

A second implication of the theory is that, even in countries in which an existing production base is in place, a surge in demand, to the extent that it attracts new entrants or new additions to capacity, may induce dumping. To understand this in the context of the model, suppose that, at the beginning of period 1, a fraction δ of the N potential country-2 z-producers know their productivity in the z sector. It is straightforward to show that if

$$(42) \quad \delta \leq \pi(1-\varphi)/(1-\pi\varphi-\pi)$$

Proposition 10 implies that country 2 will

dump z in country 1, despite the fact that country 2 has in place an existing base of producers who share with producers in country 1 a common stock of knowledge about the technology for making z.[25] Among the cases I investigated in which dumping was associated with a surge in demand for the product and a concomitant substantial expansion in output, I will mention two: the dumping of steel by Brazil in 1983 and 1984 and the dumping of color television sets by Taiwan in 1983.

In October 1982, just before the end of the 1981–1982 recession, the United States placed quota limits on the steel exports of the European Economic Community and reached an informal agreement with the Japanese to limit their U.S. exports (I. M. Destler, 1986 p. 130). In 1983, as the U.S. economy began to recover and "in the aftermath of the quota limits placed on EEC steel in October 1982 [and the restraint demonstrated by Japanese producers]... a large upsurge of developing country steel...flowed into the United States" (John Conybeare, 1987 p. 212). A rash of dumping complaints, especially against Brazilian steel exporters, followed. According to the data compiled by Destler (1986), no fewer than eight findings of dumping steel at less than fair value were issued by the Commerce Department against Brazil in 1983 and 1984. To name just two, the department, in response to complaints filed by the Bethlehem Steel and United States Steel corporations, ruled that at least four Brazilian producers were dumping hot rolled steel plate and hot rolled steel sheet in the United States (*Federal Register*, Vol. 48 [1983] p. 40419). Because of insufficient domestic sales, the Commerce Department ruled that the U.S. price did not cover the constructed value of the Brazilian steel.

The surge in steel exports from Brazil did not reflect a diversion of a relatively fixed supply to the U.S. market. According to U.S. Iron and Steel Institute data, exports of Brazilian steel to the United States were 432,000 net tons in 1979, 604,000 tons in 1982, 1,257,000 tons in 1983, 1,460,000 tons in 1984, 1,701,000 tons in 1985, and after a quota agreement between Brazil and the United States entered into effect, only 1,103,000 tons in 1986. Total production of steel by Brazil was 15,300,000 net tons in 1979, 14,324,000 tons in 1982, 16,159,000 tons in 1983, 20,266,000 tons in 1984, 22,546,000 tons in 1985, and 23,405,000 tons in 1986. Thus, between 1982 and 1985, the near trebling of Brazil steel exports to the United States accounted for less than 12 percent of the increase in total steel output, which rose by nearly 60 percent.[26] Furthermore, this was not a period during which the total U.S. demand for steel was falling. In fact total U.S. steel demand (imports plus domestic production less exports) was roughly 80,000,000 tons in 1982, 85,000,000 tons in 1983, 100,000,000 tons in 1984, and 96,000,000 tons in 1985.[27]

Consider next the dumping of color television sets by Taiwan. In 1983, in response to a complaint filed by six U.S. firms, the Commerce Department ruled that Taiwan producers were dumping color television sets in the United States. In its ruling (*Federal Register*, Vol. 48 [1983] p. 48491), the Commerce Department noted that, because at least two firms, RCA and Hitachi, had no sales of color television sets in Taiwan or in third markets, the finding of dumping was based upon sales in the United States at a price that fell below constructed value; price discrimination was of course neither alleged nor established for these firms. Total U.S. demand for television sets rose by more than 20 percent between 1982

[25]To see this, note that $p < \omega$ if $\alpha M(\omega + w)/\omega < \pi(1+\varphi)N + \pi(1+\varphi)\delta N + (1-\delta)N$. When (42) is satisfied, $[\pi(1+\varphi) + \pi(1+\varphi)\delta + (1-\delta)] > \pi(1+\varphi) + 1 - \pi(1-\varphi)$, so that (40) is sufficient for dumping to occur.

[26]I was unable to obtain plant-level data on entry and exit in the Brazilian steel industry. However, according to the industry publication *American Metal Markets*, Brazil continued to add to its steel capacity during the mid-1980's while world steel capacity was shrinking (see "A Continued Shakeout Among Steelmakers Is Probable Through the Early 1990's" [*American Metal Markets*, 1 November 1985, p. 1]).

[27]Of course, the U.S. steel industry contracted substantially between 1979 and 1982. In 1979, total U.S. demand for steel-mill products was approximately 115,000,000 tons, and total imports of steel-mill products were 17,500,000 tons.

and 1983 to more than 21,000,000 sets, while total U.S. production rose by more than 10 percent, to more than 11,000,000 sets.[28] Total exports of color television sets from Taiwan rose from 600,000 sets in 1982 to 1,500,000 sets in 1983. This did not represent a diversion of a given output level. According to industry trade publications, total output of color television sets in Taiwan was less than 1,300,000 units in 1980; by 1984 total exports alone were approaching 2,000,000 sets.[29]

A third implication of the model is that dumping should be observed in countries whose firms have not yet mastered an existing technology or have not yet acquired sufficient experience in producing the dumped good. An examination of the record reveals several instances in which countries have been found dumping goods for which there is no home market in the United States during a cyclical expansion. Such goods would not appear to be capital- or technology-intensive. Two will be mentioned here: the dumping of photograph albums by Hong Kong in 1984 and the dumping of thermostatically controlled appliance plugs by Malaysia in 1987.

In 1985, in response to a complaint filed by Esselte Pendaflex, Inc., in February 1985, the Commerce Department ruled (*Federal Register*, Vol. 50 [1985] p. 28830) that exports of photograph albums from Hong Kong were being dumped in the United States. The department calculated the constructed value of the albums because there was no home market for the goods. In 1988, in response to a complaint filed by the Triplex Inter Control Company in May 1988, the Commerce Department ruled (*Federal Register*, Vol. 53 [1988] p. 37827) that exports of thermostatically controlled appliance plugs from Malaysia were being dumped in the United States. Again, the Commerce Department based its dumping findings on the constructed-value calculation because of insufficient or nonexistent domestic market sales; price discrimination was neither alleged nor documented. A check of seven-digit SIC data revealed that the U.S. production of photograph albums (SIC 27824 41) actually rose substantially between 1983 and 1984 and was essentially unchanged in 1985, while the U.S. production of "temperature responsive controls for major appliances" (SIC 38220 75) was essentially flat in 1985, 1986, and 1987.

It should be noted that, without rather detailed data on entry and exit at the plant level in the dumping country, it is not formally possible to confirm the theory of entry, dumping, and shakeout presented in this paper. Only very recently have researchers begun to characterize actual patterns of firm entry and exit in the U.S. economy. As pointed out recently by Timothy Dunne et al. (1988 p. 45), "Despite wide recognition of the important role of entry and exit, surprisingly little is known about the patterns of firm entry and exit in the US economy." Nonetheless, the model presented here does help to provide one explanation for dumping cases in which (i) price discrimination is neither alleged nor documented, (ii) there is a surge in total demand for the dumped product, (iii) there is a surge in total production of the dumped product in the dumping country, and (iv) the dumping country is among the group of newly industrializing countries whose share of world manufacturing trade has risen so damatically during the past ten years.

VI. Concluding Remarks

This paper has developed a model of dumping that is similar in spirit to Jovanovic (1982), but in which entry and exit and the terms of trade are endogenous. The novel feature of the two-country Ricardian model developed in this paper is that dumping can arise in the free-trade equilibrium even if the two countries differ only in their initial stock of technological knowledge.

[28]Although the Census Bureau reports import volumes separately for color and black-and-white sets, it only reports domestic production by five-digit SIC code 36512, "television receivers."

[29]For production and export data on Taiwan color television sets, see *The Japan Economic Journal* (21 June 1981, p. 9), *The Journal of the Asian Economics Union* (June 1984, p. 16), and *Les Echos* (28 November 1983, p. 9).

Intuitively, dumping can arise because the present value of entry includes the value to the individual producer of learning his productivity parameter. Such information is valuable because the most productive producers earn positive rents in equilibrium. Competition among *ex ante* identical producers for these present and future rents can drive down the price of the product sufficiently so that the industry suffers losses. *Ex post* inefficient producers, upon learning their productivity parameter, recognize their losses and exit the sector. The rents earned by the most productive producers offset the opportunity cost previously incurred.

The reason why entry to the z sector is costly for a new entrant is that, for the length of a "production period," that entrant is unable to produce and consume w units of y. In this sense, factor specificity is crucial for this, as well as other, explanations of dumping. However, a specific-factors explanation that excludes a role for learning via production would not appear to predict that dumping should be observed in periods of high demand for the dumped product. Specific-factor models of the sort developed by Ethier (1982) illustrate why dumping can occur when demand is transitorily low, not permanently high.

In my analysis, I have said that a firm dumps if it exports at a price that fails to cover the opportunity cost w. Of course, with the simple Ricardian-entrepreneur setup of this paper, there is no explicit wage bill paid by a firm. Rather, there is a return that must be earned if the Ricardian entrepreneur is to remain in the z sector. Given the Commerce Department's method of constructed value, the use of the static opportunity cost of resources in this world (w) to define dumping does not seem to be at variance with the definition that is actually employed by the Commerce Department in dumping cases.

In my analysis, I have abstracted from the durability of dumped goods. If dumped goods can be purchased and stored, it is natural to ask why they are sold when the price is expected to increase, because of the mechanism outlined here, because of the mechanism outlined in Ethier (1982), or because of the expectation that the Commerce Department will impose dumping duties on the culprits. Does this not violate the intertemporal arbitrage condition? In the absence of transaction costs, it of course does. In fact, many if not most of the goods that are said by the Commerce Department to be dumped are durable and storable. The fact that I, along with others, ignore this question does not suggest that it is not important.

In the Introduction, I alluded to the literature on cyclical dumping. Two recent papers, by Gruenspecht (1988) and Andrew Dick (1988) study the connections between internal scale economies (along a learning curve in the sense of Michael Spence [1981]), an exogenous duopoly market structure, and the resulting Nash equilibrium. The basic message of both papers is that, since the relevant marginal cost of an extra unit of output includes the reduction in future costs that will be forthcoming as a firm moves down the learning curve, observed unit cost can exceed the shadow marginal cost and, depending upon the elasticity of demand, the current price of output. Thus, dumping can result because measured unit cost overstates true marginal cost.

Clearly, the common definition of dumping embodies several related but distinct phenomena. The purpose of this paper has most surely not been to argue that all dumping is the result of the process of entry and shakeout that has been outlined above. Rather, the modest goal has been to show one mechanism by which transitory differences in the stock of technological knowledge in otherwise identical countries can result in international trade and, under certain conditions, dumping, and to evaluate some of the implications.

Appendix

PROOF OF PROPOSITION 5:

From Corollary 2, when $p' < w/(1-\varphi)$, $e < \beta$. The minimum supply of z in period 2 at a price $p' > w$ is given by

$$N[1-\pi(1-\varphi)].$$

This is because at most πN firms will be unwilling to produce at $p' > w$. These are the (at most) πN firms that entered in period 1 and found their productivity parameter to be $1-\varphi$. Since total output of z is $NE\xi = 1$, $N[1-\pi(1-\varphi)]$ is the minimum period-2 output of z at $p' \geq w$. Thus, from (13), $\alpha M(1-\varphi) \leq N[1-\pi(1-\varphi)]$ is the restriction that insures that $p' \leq w/(1-\varphi)$. Equation (17) and the second term in the max$[\cdot]$ in (18) follow immediately. For $p < w$, demand at a price w must fall short of the elastic supply of z, N, when p equals or exceeds the entry price $e < w$. This will be true when $\alpha Mw/w < N$. The first term in the max$[\cdot]$ in (18) follows. By similar reasoning, $\alpha M \leq N[1-\pi(1-\varphi)]$ is the condition that, along with the condition $\varphi \leq \frac{1}{2}$ (Corollary 2), $p' = w$. When the period-2 price is pushed to the break-even price w, the period-1 *ex ante* entry price e is, from equation (14), equal to $w(1-\pi\varphi)$.

PROOF OF PROPOSITION 7:

The proof is similar to that of Proposition 5. From Corollary 2, when $p' < w/(1-\varphi)$, $e < \beta$. The minimum world supply of z in period 2 at a price $p' \geq w$ is given by $2N[1-\pi(1-\varphi)]$. This is because at most πN firms in each country will be unwilling to produce at $p' > w$. These are the (at most) πN country-2 firms that entered in period 1 and found their productivity parameter to be $1-\varphi$, plus the πN inefficient country-1 firms. Since total output of z is $2NE\xi = 2N$, $2N[1-\pi(1-\varphi)]$ is the minimum period-2 output of z at $p' \geq w$. Thus, from (13), $2\alpha M(1-\varphi) \leq 2N[1-\pi(1-\varphi)]$ is the restriction that insures that $p' \leq w/(1-\varphi)$. The second term in max$[\cdot]$ in (23) follows immediately. For the first-period equilibrium price to be driven below the break-even price, $p < w$, world demand at a price w must fall short of the elastic supply of z when p equals or exceeds the entry price $e < w$. This supply is given by $N + \pi(1+\varphi)N$. It is the sum of elastic country-2 supply at the entry price e and the supply of the most efficient producers in country 1. When $\alpha Mw/w < N + \pi(1+\varphi)N$, $p < w$ if $p' < w/(1-\varphi)$. The first term in max$[\cdot]$ in (23) follows. Finally, for dumping to occur in period 1, it must be the case that $d_2(p) > 2\pi(1+\varphi)N$; otherwise country 1 will export z to country 2. Since $p < w$ by the first inequality, this condition will be satisfied if $2\alpha Mw/w > 2\pi(1+\varphi)N$; the second inequality follows. The set of N that satisfies (24) is nonempty for all $\pi < \frac{1}{2}$, $\varphi < 1$, and positive α, since $[1-\pi(1-\varphi)] > \pi(1+\varphi)$ and $1+\pi(1+\varphi) > 2\pi(1+\varphi)$.

REFERENCES

Alchian, Armen, "Uncertainty, Evolution, and Economic Theory," *Journal of Political Economy*, June 1950, *58*, 211–21.

Bhagwati, Jagdish, *Protectionism*, London: Blackwell, 1986.

Conybeare, John, *Trade Wars: The Theory and Practice of International Commercial Rivalry*, New York: Columbia University Press, 1987.

Destler, I. M., *America Trade Politics*, Washington, DC: Institute for International Economics, 1986.

Dick, Andrew, "Imperfect International Competition in Innovation and Product Markets: Theoretical Implications and Empirical Results," mimeo, University of Chicago, 1988.

Dixit, Avinash and Shapiro, Carl, "Entry Dynamics with Mixed Strategies," in Lacy Thomas, ed., *The Economics of Strategic Planning*, Boston: Lexington Books, 1986, pp. 63–79.

Dunne, Timothy, Roberts, Mark J. and Samuelson, Larry, "Patterns of Firm Entry and Exit in U.S. Manufacturing Industries," *Rand Journal of Economics*, Winter 1988, *19*, 495–515.

Ethier, Wilfred, "Dumping," *Journal of Political Economy*, June 1982, *90*, 487–506.

Gort, Michael and Klepper, Steven, "Time Paths in the Diffusion of Product Innovations," *Economic Journal*, September 1982, *92*, 630–53.

Gruenspecht, Howard, "Dumping and Dynamic Competition," *Journal of International Economics*, November 1988, *25*, 225–48.

Jovanovic, Boyan, "Selection and the Evolution of Industry," *Econometrica*, May 1982, *50*, 649–70.

_____ **and Lach, Saul,** "Entry, Exit, and Diffusion with Learning by Doing," *American Economic Review*, September 1989, *79*, 680–99.

Hopenhayn, Hugo, "A Competitive Stochastic Model of Entry and Exit to Industry," mimeo, University of Minnesota, 1986.

Lippman, Steven and Rumelt, R., "Uncertain Imitability: Interfirm Differences in Efficiency Under Competition," *Bell Journal of Economics*, Autumn 1982, *13*, 418–38.

Nelson, Richard R. and Winter, Sidney G., *An Evolutionary Theory of Economic Change*, Cambridge, MA: Harvard University Press, 1982.

Olley, Steven and Pakes, Ariel, "The Dynamics of Productivity in the U.S. Telecommunications Industry," mimeo, Yale University, 1988.

Pakes, Ariel and Ericson, Richard, "An Alternative Model of Firm and Industry Dynamics," mimeo, Yale University, 1988.

Spence, A. Michael, "The Learning Curve and Competition," *Bell Journal of Economics*, Spring 1981, *12*, 49–70.

Viner, Jacob, *Dumping*, Chicago: University of Chicago Press, 1923.

Economic Report of the President, Council of Economic Advisers, Washington, DC: U.S. Government Printing Office, 1988.

Annual Report on the Operation of the Trade Agreement Program, U.S. International Trade Commission, Washington, DC: U.S. Government Printing Office, various years.

Union Organizing Activity, Firm Growth, and the Business Cycle

By STEPHEN G. BRONARS AND DONALD R. DEERE*

This paper analyzes the relationship between changes in unionization and firm growth. Average growth is significantly lower in manufacturing firms that experience successful union elections, but these strong "effects" are largely illusory. We find no evidence of a significant relationship between unionization and firm growth, despite a strong cyclical pattern in election activity. Our results suggest that the significant negative effect of organizing activity on a firm's market value is not accompanied by any growth changes. We therefore cannot reject the hypothesis that the equity losses from union election activity represent a simple transfer of wealth from shareholders to workers. (*JEL* J50)

A large body of empirical evidence indicates that unionization is negatively related to firm profitability (see e.g., Kim B. Clark, 1984; Michael A. Salinger, 1984; Thomas Karier, 1985; Robert A. Connolly et al., 1986; Paula B. Voos and Lawrence R. Mishel, 1986; Nancy L. Rose, 1987; John M. Abowd, 1989a; Barry T. Hirsch, 1991).[1] In particular, increases in union organizing activity at a firm result in significant losses for shareholders. Richard S. Ruback and Martin B. Zimmerman (1984) and Bronars and Deere (1990) find that the market value of the firm declines by 3–4 percent following a successful union representation election and by 1–2 percent after an unsuccessful election. These estimates surely understate the impact of unionization on profitability because stock prices react only to new information. Given the sizable effect of union organizing activity on the market values of firms, it is important also to examine the relationship between union representation elections and firms' employment, investment, and output decisions.

If firms alter their inputs or output in response to a union election, then a portion of the associated decline in equity represents a loss to shareholders that the union does not capture. On the other hand, if unions extract wealth efficiently, a union election will not cause changes in firms' behavior but will merely transfer income to union members.[2]

Unions achieve wage gains by appropriating rents or quasi-rents that would otherwise accrue to shareholders. Increases in employment, capital investment, or output generate such rents and may thus increase union organizing activity. Alternatively,

*Bronars: Department of Economics, University of Texas, Austin, TX 78712; Deere: Department of Economics, Texas A&M University, College Station, TX 77843. A portion of the work on this paper was completed while Deere was on leave at UC-Santa Barbara. This material is based in part upon work supported by the Texas Advanced Research Program under grant number 14581 to Deere. We thank George Borjas, Melissa Famulari, Jeff Grogger, Brooks Pierce, Robert Reed, Doug Steigerwald, Larry Summers, Joe Tracy, Steve Trejo, and an anonymous referee for helpful comments. We are grateful to Richard Ruback and Martin Zimmerman for generously providing us with their data. Peter Griffin provided excellent research assistance. Any remaining errors are ours.

[1]In isolated cases higher union productivity may offset the union wage gains (see e.g., Charles Brown and James L. Medoff, 1978; Kim B. Clark, 1980; Steven G. Allen, 1984, 1987).

[2]Analyzing firms' behavior around union elections thus provides a useful method for gauging the efficiency of new unionization. There have been several recent analyses of the efficiency of union contracting between a firm and an existing union, with mixed results (see Clark, 1984; James N. Brown and Orley C. Ashenfelter, 1986; David Card, 1986; Randall W. Eberts and Joe A. Stone, 1986; Thomas E. MaCurdy and John H. Pencavel, 1986; Abowd, 1989a).

growing firms may be less susceptible to unionization because of an increased ability to maintain worker morale. A careful analysis of the changes in inputs and output and the occurrence of union elections can help to clarify the causal relationship between firm behavior and union organizing activity.

This paper examines how union representation elections and employment, investment, and output decisions are related using a 20-year panel of data on large publicly traded firms. Average growth rates of inputs and output are indeed significantly lower in firms that experienced successful union elections in the manufacturing sector, but these apparently strong "effects" of unionization are largely illusory.

In our search for causal linkages, we perform vector autoregressions of union elections and firm growth. Statistical tests for Granger-causality provide some evidence of causation running from union elections to firm growth, but no indication of causation from growth to elections. The direction of the effects on growth differ across manufacturing and nonmanufacturing firms, however, and the strongest effects on growth are associated with elections that the union loses, while there is essentially no change in growth after union wins. We do find that representation elections tend to occur near peaks in aggregate economic activity, making it somewhat difficult to isolate the effects of these elections from cyclical factors.

Taken together these results provide the key finding of the paper: that while union election activity has a strong cyclical component, it appears to be largely unrelated to changes in the firm's inputs or output. Thus, the main implication of the analysis is that the negative equity effect of union elections is not accompanied by any impact on firm growth.

The next section describes the data set that allows us to test for causality between union organizing activity and a firm's input and output decisions and briefly summarizes the important empirical regularities in union election activity and firm growth. This section documents the significant, negative cross-section relationship between elections and growth and highlights the strong cyclical pattern in election activity. Section II presents the results of the vector autoregressions and causality tests. The paper concludes with a discussion of our findings and their implications.

I. The Data

The events that we use to signify changes in union status at a firm are similar to those used by Ruback and Zimmerman (1984). These data are based on union representation elections filed with the National Labor Relations Board (NLRB). The first step in the unionization process is to petition the NLRB to hold an election. A petition is successful if it contains signatures of at least 30 percent of the employees to be covered by the new union. Successful petitions confirm that an election will take place, usually within several months of the filing of the petition. Election results are certified by the NLRB, typically within five days of the actual election. Following Ruback and Zimmerman (1984), we initially limit our sample to those elections involving 750 or more workers in NYSE firms between February 1962 and December 1984.

We match these data on union representation elections to observations on firms' behavior obtained from Standard and Poor's COMPUSTAT data. These annual data, covering the period 1965–1984, are derived from the financial statements, 10-K reports, and other SEC filings of publicly traded firms and represent year-end values. We focus on the time-series behavior of a firm's employment, investment in plant and equipment, and sales revenue.[3] We deleted firms from our sample if they did not report employment, investment, or sales for any of the years from 1965 to 1984.

There are a total of 85 firms for which we can match union election data (involving at least 750 workers) and a complete 20-year panel of employment, investment, and sales data from COMPUSTAT. The merging of these two data sets results in a sample of 229 elections. Table 1 presents descriptive

[3]Sales and investment are measured in millions of 1972 dollars, and employment is measured in thousands of workers. These values represent year-end totals/levels.

TABLE 1—UNION REPRESENTATION ELECTIONS 1962–1984 INVOLVING AT LEAST 750 WORKERS

A. *Distribution of Elections Across Firms*

Elections per firm	Number of firms	Percentage of firms	Union wins per firm	Number of firms	Percentage of firms
1	44	51.76	0	49	57.65
2	14	16.47	1	27	31.76
3	8	9.41	2	5	5.88
4	4	4.71	3	2	2.35
5	5	5.88	4	2	2.35
6	2	2.35			
7	2	2.35			
8	2	2.35			
9	3	3.53			
23	1	1.18			
Total:	85	100.00		85	100.00

B. *Distribution of Elections Through Time*

Year	Number of elections	Number of wins
1962	9	2
1963	11	2
1964	9	2
1965	11	1
1966	7	2
1967	11	5
1968	18	4
1969	13	3
1970	11	4
1971	14	2
1972	9	2
1973	22	4
1974	21	2
1975	8	1
1976	3	0
1977	5	2
1978	10	3
1979	14	1
1980	10	4
1981	3	0
1982	5	3
1983	1	0
1984	4	2
Total:	229	51

C. *Size Distribution of Elections*

	All elections		Successful elections	
Percentile	Number of workers involved	Percentage of workers involved	Number of workers involved	Percentage of workers involved
1	800	0.2	750	0.1
25	1,035	0.9	995	0.7
50	1,210	2.5	1,314	2.6
75	1,717	5.2	1,729	5.2
100	8,408	74.6	5,525	74.6
N:	229	200	51	45
Mean:	1,543	5.2	1,538	5.3
SD:	965	9.0	864	11.3

TABLE 2—SUMMARY STATISTICS FOR ELECTION AND NONELECTION FIRMS, 1965–1984

Variable	Firms where the union won an election	Firms where the union did not win	Nonelection firms
Number of firms	36	49	72
Number of elections	128	101	0
Number of wins	51	0	0
Percentage of firms in manufacturing	69	90	75
Sample means (standard deviations):			
Sales in $ millions	3,563.5 (6,240.9)	1,547.0 (1,661.9)	2,951.7 (5,258.0)
Investment in $ millions	261.7 (492.5)	107.1 (145.4)	246.1 (510.0)
Employment in 1,000's	98.1 (151.6)	42.4 (39.2)	56.4 (57.7)
Log sales	7.38 (1.20)	6.76 (1.22)	7.20 (1.28)
Log investment	4.66 (1.31)	3.89 (1.41)	4.36 (1.60)
Log employment	3.84 (1.22)	3.21 (1.18)	3.48 (1.19)

statistics for the sample of union elections, including information on the distribution of elections across firms and through time and on the amount of employment involved in these elections. The table shows that the fraction of employment involved in these elections is skewed, though the distributions are very similar for all elections and for elections that the union wins.

In order to obtain a control group with which to compare our sample of union-election firms, we selected firms traded on the New York Stock Exchange that did not experience a union representation election involving 750 or more workers in the sample period. We again deleted all firms for which there were missing data concerning employment, investment, or sales in the COMPUSTAT data set. To limit the impact of differences in firm size across election and control groups a stratified sampling procedure was used. We first calculated the sample quartiles of average employment over the period 1965–1984 for the 85 election firms we observed. Using these sample quartiles we divided the nonelection firms into four mutually exclusive groups based on their average employment during 1965–1984. We then obtained the largest possible random sample of nonelection firms given this stratification.[4] The resulting control group includes 72 nonelection firms.

Table 2 presents summary statistics for the election and nonelection subsamples. The election subsample is divided into those firms that experienced at least one successful election and those firms that did not. The industry composition of the nonelection and winning-election firms is nearly identical, though losing-election firms are

[4]The sample size of nonelection firms was based on the election-firm quartile that contained the smallest number of firms: there were 18 nonelection firms with average employment in the top quartile. After including all of these firms in our nonelection sample, we randomly sampled firms from the other quartiles to yield equal numbers of firms in each employment size group. It is noteworthy that the employment size distribution of potential control firms (i.e., no election, at least 750 workers, all 20 years of data) is very skewed compared to the distribution of election firms: there are only 18 firms in the top quartile and 408 in the bottom quartile.

TABLE 3—AVERAGE ANNUAL GROWTH RATES BY INDUSTRY FOR ELECTION AND NONELECTION FIRMS, 1965–1984 (PERCENTAGES; STANDARD DEVIATIONS IN PARENTHESES)

Variable	Firms with a union win	Firms without a union win	Nonelection firms
Manufacturing:			
Growth in sales	2.64[a]	3.77	4.82[c]
	(2.64)	(3.28)	(3.13)
Growth in investment	1.79[a,b]	4.04[c]	5.19[c]
	(3.87)	(4.55)	(5.38)
Growth in employment	0.81[a]	2.21	2.94[c]
	(2.86)	(3.35)	(3.55)
Nonmanufacturing:			
Growth in sales	4.96	4.07	5.78
	(4.39)	(3.87)	(3.11)
Growth in investment	5.80	5.58	5.44
	(5.92)	(4.98)	(4.82)
Growth in employment	3.42	3.04	4.27
	(4.21)	(5.21)	(3.44)

[a]Difference from the growth rate for nonelection firms is significantly different from zero at the 0.05 level.
[b]Difference from the growth rate for firms without a union win is significantly different from zero at the 0.05 level.
[c]Difference from the growth rate for firms with a union win is significantly different from zero at the 0.05 level.

more likely to be in manufacturing. The average winning-election firm is larger than the average nonelection firm, which is larger than the average losing-election firm. Thus firms in the control group are roughly similar to the election firms in terms of industry composition and size.

Table 3 highlights cross-section differences in the growth rates of employment, sales, and investment between winning-election, losing-election, and nonelection firms over the sample period. Within manufacturing, average growth rates in employment, investment, and sales are about 28–33 percent smaller in losing-election firms than in the control group of nonelection firms. The growth-rate differences between winning-election and nonelection manufacturing firms are even more pronounced. Employment, investment, and sales growth rates are only about one-third to one-half as large in winning-election firms as they are in the control group of nonelection firms in the manufacturing sector.[5]

[5]These results continue to hold if average growth rates are regressed on the size of the firm in 1965 and dummy variables for winning-election status, losing-election status, and durable manufacturing. Though firm size is an important (negative) determinant of growth and larger firms are more likely to experience a union win, winning-election firms still experienced significantly less growth over this period when we controlled for initial firm size.

Three recent papers also report a negative cross-section relationship between unionization and employment. David G. Blanchflower et al. (1991) analyze data on British establishments and conclude that union status is negatively related to employment growth, though Stephen J. Machin and Sushil B. Wadhwani (1991) dispute this. Jonathan S. Leonard (1992) finds that employment growth is lower by about 4 percentage points in union plants than in nonunion plants. Edward B. Montgomery (1989) uses CPS data and finds a very weak relationship between employment status and unionization.

TABLE 4—AVERAGE GNP GROWTH AND AVERAGE ELECTION-FIRM SALES GROWTH AROUND UNION ELECTIONS (PERCENTAGES; STANDARD ERRORS IN PARENTHESES)

Time	GNP growth: Observed elections	GNP growth: If elections were uniformly distributed over time	GNP growth: Difference	Average sales growth at election firms
Two years before an election	3.57	2.83	0.74** (0.18)	4.35
One year before an election	3.52	2.80	0.72** (0.18)	4.01
Year of an election	2.75	2.85	−0.10 (0.18)	4.14
One year after an election	2.10	2.74	−0.64** (0.18)	1.65
Two years after an election	2.14	2.60	−0.46** (0.17)	2.76

Note: The F test that the first two columns are equal can be rejected at the 0.01 level. See the text for further details.

**Significantly different from zero at the 0.01 level.

This negative cross-section relationship between average growth rates and union elections is not apparent in nonmanufacturing firms. In general the differences in growth rates across nonelection, losing-election, and winning-election firms are relatively minor and are always statistically insignificant.

Two interesting features of the time-series behavior of union organizing activity are apparent from Table 1. First, the probability of observing a union election decreases sharply over time. Second, representation-election propensities appear to be highest near the peak of the business cycle, 1–2 years before major cyclical contractions. Table 4 provides more direct evidence of this cyclical pattern. For each election in our sample, we calculate annual GNP growth for the year of the election and for the two years immediately before and immediately after the election.

The first column of Table 4 presents averages of these five GNP growth rates across the 200 elections that occurred during 1965–1984. GNP growth was noticeably higher on average in the years just before an election than in the years just after an election. For comparison, the average GNP growth rates that would arise if these elections were distributed uniformly over 1965–1984 are presented in the second column of Table 4. The third column reports the difference between the first two columns with associated standard errors in order to test the null hypothesis that these large union elections are uncorrelated with the business cycle.[6] The results indicate that GNP growth was significantly above average before elections and was significantly below average following elections.[7]

[6]Given that elections became increasingly rare over time, we also compared the observed GNP growth rates around elections to those that would arise if elections were distributed along a simple negative trend. Again we could reject this hypothesis in favor of a procyclical pattern.

[7]Orley C. Ashenfelter and John H. Pencavel (1969) argue that organizing is generally procyclical. As employment expands, the pool of unorganized workers increases, and representation elections are more likely to occur. In addition, union organizing drives may be more likely during expansions because there is both more to be won (i.e., profits are higher) and less to be lost (i.e., labor markets are tighter). Interestingly, the results here and in Tables 5 and 6 suggest that aggregate growth has a much stronger effect on organizing activity at a firm than does the firm's own growth.

The fourth column of Table 4 presents average sales growth in the election firms for the corresponding years surrounding an election.[8] Similar to the pattern of growth in aggregate output, growth in the firms that experience elections is higher before and lower after an election. Thus, observed differences in firm behavior before and after union elections may largely reflect effects of the business cycle, rather than the impact of election activity.

The simple cross-section and time-series relationships depicted in Tables 3 and 4 imply that: (i) average growth rates between 1965 and 1984 were significantly negatively related to union election success, at least in the manufacturing sector; and (ii) union election propensities were the highest at the peak of cyclical expansions. In the next section we develop a more formal empirical analysis with these facts in mind.

II. Vector Autoregressions of Union Elections and Firm Growth

The cross-section results described above may simply be due to unobserved heterogeneity in growth rates that is correlated with the propensity for union elections. In this section, we develop and estimate a dynamic model of union elections and firm growth that allows for firm-specific heterogeneity in growth rates and election propensities. Given the strong cyclical pattern of election activity, we control for aggregate growth by allowing separate year effects as well as by including the nonelection firms in the regressions. We then perform hypothesis tests on these vector autoregressions (VAR's) in an attempt to uncover whether increases in firm growth influence union organizing activity and whether union elections affect subsequent firm growth.

Consider the following cross-section/time-series model of a firm's input and output decisions and union organizing activity at the firm:

$$(1)\quad Y_{it} = \alpha_i + \theta_t + \sum_{m=1}^{M} \boldsymbol{\beta}_m \mathbf{Y}_{it-m} + \sum_{m=1}^{M} \boldsymbol{\gamma}_m \mathbf{X}_{it-m} + \varepsilon_{it}$$

where Y_{it} represents the annual percentage change (log difference) in firm i's sales, employment, or investment in plant and equipment. The coefficients $\boldsymbol{\beta}_m$ measure the impact of lagged changes in the growth of inputs or output on the current changes in inputs or output. The vector $\mathbf{X}_{it}$ contains two variables measuring the union election activity at firm i in period t: the fraction of the firm's employment involved in unsuccessful union elections and the fraction in successful elections.[9] The coefficients $\boldsymbol{\gamma}_m$ measure the impact of past union organizing activity on the time-series pattern of sales, employment, or investment at the firm. The parameters α_i allow for differences in the mean growth rates of sales, employment, or investment across firms, while the θ_t allow for separate year effects. The lag length, equal to M, is sufficient to capture all of the impact of lagged Y and X on current Y; hence, the error term ε_{it} is a zero-mean serially uncorrelated random variable.

[8]Average sales growth includes only those firm/years that were within two years of an election. This can be most easily calculated by regressing the 1,615 election-firm observations (19 years [one is lost in calculating growth rates]×85 firms) on a constant, an indicator variable for the occurrence of an election, and two leads and two lags of this indicator. The patterns of sales growth and GNP growth are quite similar to those reported if the sample is divided into manufacturing-sector and nonmanufacturing-sector elections. There are also similar patterns of growth in employment and capital investment in the election firms.

[9]We also performed the analysis using simple dummy variables as the measure of union election activity. These results were similar to (though somewhat weaker than) those reported. In Bronars and Deere (1990), we found that the fraction of firm employment involved in the election is an important determinant of the equity loss associated with the election; hence, we prefer the use of the employment fractions.

The time-series behavior of $\mathbf{X}_{it}$ can be modeled in a similar fashion:

$$(2)\quad X_{it} = \kappa_i + \lambda_t + \sum_{m=1}^{M} \boldsymbol{\delta}_m \mathbf{X}_{it-m} + \sum_{m=1}^{M} \boldsymbol{\pi}_m \mathbf{Y}_{it-m} + \nu_{it}$$

where the coefficients $\boldsymbol{\delta}_m$ measure the impact of lagged union elections on current elections, and the coefficients $\boldsymbol{\pi}_m$ measure the impact of lagged sales, employment, or investment on union elections. The parameters κ_i and λ_t allow union election activity to differ across firms and over time, respectively. Again the lag length M ensures that the error term ν_{it} is serially uncorrelated. Analysis of the estimated coefficient vectors, $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$, provides the opportunity to test for Granger-causality between elections and growth.

Equations (1) and (2) specify a complete dynamic relationship between elections and firm growth. The structure of this relationship may, however, differ across groups of firms. For example, as was evident in the preceding section, the relationship between union election activity and average firm growth appears to differ across major industry groups. In view of this, we estimate equations (1) and (2) separately for manufacturing and nonmanufacturing firms.[10]

It is important to note that ordinary least-squares (OLS) estimation of (1) and (2) may lead to biased coefficient estimates because of the presence of lagged endogenous variables (Stephen J. Nickell, 1981). Given that our panel is rather long, such biases may be small. An alternative estimation strategy, which eliminates this bias via a generalized two-stage least-squares (2SLS-GLS) procedure, is suggested by Douglas Holtz-Eakin et al. (1988). We performed both estimation procedures. A comparison of the OLS and 2SLS-GLS results reveals that the OLS estimates of the own-lag effects do differ significantly, but predictably, from the 2SLS-GLS estimates. Hence the Nickell (1981) bias is present and is roughly equal to $-1/(T-1)$, where $T = 16$ is the effective length of the panel after lagging. Despite these differences in own-lag effects, the estimates of the cross-effects are quite similar. The implications for causality are therefore the same using either estimation procedure. Given this comparison, we report the 2SLS-GLS results in Tables 5 and 6. For completeness, the OLS results are given in the Appendix, along with a discussion of the data sources and methodology. We turn now to a discussion of the results.

The estimates of (1) reported in Table 5 provide some evidence that union election activity causes the firm's subsequent growth to change. Employment and sales growth decline following a losing election in manufacturing. An election involving 3 percent of the firm's workers decreases growth by about 0.5–0.75 percentage points in the next three years. This effect is somewhat stronger and of greater statistical significance for sales growth. In contrast to the negative effect of

[10]There are 69 election firms (with 180 elections) in manufacturing and 16 election firms (with 49 elections) in nonmanufacturing. There are 54 manufacturing and 18 nonmanufacturing firms in the nonelection control group.

We also estimated a variety of other specifications. Pooling all firms across industries gave no evidence of any causal relationships. If those firms that experienced only a single election (35 in manufacturing and nine in nonmanufacturing) were included along with the nonelection firms, the results are similar to and slightly stronger than those reported. The variation in GNP growth around these single elections is also more pronounced than that reported in Table 4, however. We estimated VAR's that included lagged sales, investment, and employment growth together as regressors. These results were in close agreement with those reported. Lastly, because the vast majority of elections involve fewer than 750 workers, we also examined the complete union-election histories for the firms in our sample over the period 1977–1984 (the NLRB data prior to 1977 are very difficult if not impossible to match to COMPUSTAT data). These 157 firms experienced 1,541 elections during 1977–1984 of which 642 were successful for the union. Of the 72 firms originally designated nonelection, 52 experienced an election involving fewer than 750 workers during 1977–1984. We estimated VAR's similar to those reported in Tables 5 and 6, and also measured the election effects with simple dummy variables. These additional estimates were generally supportive of our findings: the estimated effects varied across industries, and there was little evidence of union-win effects.

TABLE 5—VECTOR AUTOREGRESSIONS (2SLS-GLS): GROWTH

	Dependent variable					
	Growth in employment		Growth in investment		Growth in sales	
Explanatory variables	Coefficient	SE	Coefficient	SE	Coefficient	SE
Firms in Manufacturing:						
Lag-1 union loss	−1.77	1.25	−2.85	4.39	−2.47	1.20
Lag-2 union loss	−2.64	1.25	0.53	4.40	−1.78	1.20
Lag-3 union loss	−1.36	1.25	−0.31	4.39	−2.06	1.72
Lag-1 union win	−2.54	5.83	13.25	20.46	−3.32	5.60
Lag-2 union win	−0.84	3.71	5.76	13.03	−1.49	3.57
Lag-3 union win	−0.76	3.53	1.65	12.38	1.28	3.39
Lag-1 employment	0.13	0.03	—	—	—	—
Lag-2 employment	0.002	0.02	—	—	—	—
Lag-3 employment	−0.04	0.02	—	—	—	—
Lag-1 investment	—	—	−0.19	0.02	—	—
Lag-2 investment	—	—	−0.23	0.02	—	—
Lag-3 investment	—	—	−0.12	0.02	—	—
Lag-1 sales	—	—	—	—	0.21	0.03
Lag-2 sales	—	—	—	—	−0.02	0.02
Lag-3 sales	—	—	—	—	−0.04	0.02
F tests (*p* values in parentheses):						
All union-loss coefficients	1.90 (0.127)		0.16 (0.921)		2.25 (0.080)	
All union-win coefficients	0.08 (0.969)		0.19 (0.903)		0.23 (0.874)	
All election coefficients	1.01 (0.419)		0.18 (0.983)		1.26 (0.271)	
Firms in Nonmanufacturing:						
Lag-1 union loss	3.33	2.32	11.10	10.09	9.23	2.79
Lag-2 union loss	3.47	2.31	18.21	10.12	9.72	2.82
Lag-3 union loss	3.42	2.28	12.03	9.99	5.33	2.80
Lag-1 union win	3.81	4.95	15.03	21.91	−2.13	6.03
Lag-2 union win	3.81	4.95	15.03	21.91	2.48	6.03
Lag-3 union win	−1.16	1.21	−9.34	5.61	1.11	1.50
Lag-1 employment	0.12	0.05	—	—	—	—
Lag-2 employment	0.08	0.05	—	—	—	—
Lag-3 employment	0.06	0.04	—	—	—	—
Lag-1 investment	—	—	−0.26	0.05	—	—
Lag-2 investment	—	—	−0.10	0.05	—	—
Lag-3 investment	—	—	−0.01	0.05	—	—
Lag-1 sales	—	—	—	—	0.17	0.05
Lag-2 sales	—	—	—	—	−0.02	0.05
Lag-3 sales	—	—	—	—	0.20	0.05
F tests (*p* values in parentheses):						
All union-loss coefficients	1.66 (0.175)		1.53 (0.206)		6.81 (0.001)	
All union-win coefficients	0.70 (0.552)		1.37 (0.252)		0.30 (0.829)	
All election coefficients	1.17 (0.319)		1.45 (0.195)		3.58 (0.002)	

Note: Each regression also included a complete set of year dummy variables. The coefficients on the union variables have been multiplied by 10. For firms in manufacturing, $N = 1{,}845$; for firms in nonmanufacturing, $N = 510$.

TABLE 6—VECTOR AUTOREGRESSIONS (2SLS-GLS): ELECTIONS
(STANDARD ERRORS IN PARENTHESES)

	Dependent variables					
	Elections lost by the union			Elections won by the union		
Explanatory variables	(i)	(ii)	(iii)	(i)	(ii)	(iii)
Firms in Manufacturing:						
Lag-1 union loss	−3.35	−3.28	−3.33	0.16	0.19	0.22
	(2.48)	(2.49)	(2.49)	(0.53)	(0.53)	(0.53)
Lag-2 union loss	−2.32	−2.25	−2.37	0.51	0.56	0.51
	(2.49)	(2.49)	(2.49)	(0.53)	(0.53)	(0.53)
Lag-3 union loss	−3.32	−3.46	−3.19	−0.06	−0.04	−0.02
	(2.49)	(2.49)	(2.49)	(0.53)	(0.53)	(0.53)
Lag-1 union win	−6.55	−6.12	−6.10	−1.40	−1.44	−1.37
	(11.60)	(11.59)	(11.59)	(2.45)	(2.45)	(2.45)
Lag-2 union win	1.93	2.00	1.71	−0.79	−0.69	−0.86
	(7.38)	(7.38)	(7.38)	(1.56)	(1.56)	(1.56)
Lag-3 union win	−1.64	−2.02	−1.56	−1.48	−1.53	−1.51
	(7.02)	(7.02)	(7.01)	(1.48)	(1.48)	(1.48)
Lag-1 employment	−0.33	—	—	−0.83	—	—
	(5.32)			(1.12)		
Lag-2 employment	−3.31	—	—	0.14	—	—
	(4.89)			(1.03)		
Lag-3 employment	−0.38	—	—	−1.02	—	—
	(4.69)			(1.99)		
Lag-1 investment	—	0.62	—	—	0.23	—
		(1.41)			(0.30)	
Lag-2 investment	—	0.03	—	—	0.27	—
		(1.41)			(0.30)	
Lag-3 investment	—	−1.90	—	—	−0.10	—
		(1.41)			(0.30)	
Lag-1 sales	—	—	0.17	—	—	−2.11
			(5.71)			(1.21)
Lag-2 sales	—	—	3.92	—	—	0.66
			(5.07)			(1.07)
Lag-3 sales	—	—	−6.13	—	—	−1.85
			(4.96)			(1.05)
F tests (*p* values in parentheses):						
All employment coefficients	0.16	—	—	0.49	—	—
	(0.926)			(0.689)		
All investment coefficients	—	0.87	—	—	0.56	—
		(0.459)			(0.643)	
All sales coefficients	—	—	0.72	—	—	1.90
			(0.543)			(0.128)

TABLE 6—*Continued*

	Dependent variables					
	Elections lost by the union			Elections won by the union		
Explanatory variables	(i)	(ii)	(iii)	(i)	(ii)	(iii)
Firms in Nonmanufacturing:						
Lag-1 union loss	1.61	1.50	1.33	0.19	0.25	0.33
	(4.87)	(4.81)	(4.83)	(0.68)	(0.68)	(0.68)
Lag-2 union loss	3.08	2.11	2.31	0.33	0.26	0.33
	(4.86)	(4.86)	(4.88)	(0.68)	(0.68)	(0.69)
Lag-3 union loss	3.63	3.77	2.86	−0.22	−0.20	−0.22
	(4.79)	(4.76)	(4.85)	(0.67)	(0.67)	(0.68)
Lag-1 union win	−1.84	−2.07	−1.80	−0.19	−0.25	−0.22
	(10.45)	(10.44)	(10.44)	(1.47)	(1.47)	(1.47)
Lag-2 union win	3.25	3.60	3.05	−0.50	−0.56	−0.47
	(10.42)	(10.45)	(10.44)	(1.46)	(1.47)	(1.47)
Lag-3 union win	0.39	0.75	0.55	0.01	−0.08	−0.03
	(2.55)	(2.67)	(2.60)	(0.36)	(0.38)	(0.37)
Lag-1 employment	−7.92	—	—	1.82	—	—
	(10.79)			(1.51)		
Lag-2 employment	−14.30	—	—	−1.18	—	—
	(9.58)			(1.34)		
Lag-3 employment	1.80	—	—	0.60	—	—
	(8.93)			(1.25)		
Lag-1 investment	—	−2.14	—	—	0.42	—
		(2.40)			(0.34)	
Lag-2 investment	—	−0.11	—	—	0.15	—
		(2.46)			(0.35)	
Lag-3 investment	—	−1.18	—	—	0.08	—
		(2.32)			(0.33)	
Lag-1 sales	—	—	1.66	—	—	−0.12
			(9.42)			(1.32)
Lag-2 sales	—	—	−4.23	—	—	−0.14
			(8.90)			(1.25)
Lag-3 sales	—	—	11.66	—	—	−0.16
			(8.80)			(1.24)
F tests (*p* values in parentheses):						
All employment coefficients	0.87	—	—	1.01	—	—
	(0.455)			(0.386)		
All investment coefficients	—	0.37	—	—	0.52	—
		(0.773)			(0.671)	
All sales coefficients	—	—	0.68	—	—	0.01
			(0.565)			(0.999)

Note: Each regression also included a complete set of year dummy variables. The coefficients on the election variables and the growth variables have been multiplied by 100 and 1,000, respectively. For firms in manufacturing, $N = 1{,}845$; for firms in nonmanufacturing, $N = 510$.

elections in manufacturing, growth increases after a losing election in nonmanufacturing firms. Growth is 1–5 percentage points higher in the three years following a losing election involving 3 percent of the firm's employees. The effect on sales growth is quite statistically significant, whereas the effects on employment and investment growth are marginally significant at best. Despite these results for losing elections, however, there is virtually no evidence that successful union elections have any effect on firm growth in either industry.[11]

The formal statistical tests for causality entail examining whether changes in union election activity help predict firm growth given past growth. The joint F tests that all three lags of the election variables equal zero are also presented in Table 4. Noncausality can be rejected in only two cases: the effect of losing elections on sales growth in each industry. There is little evidence of any other causal effects and no indication that union wins have any significant impact on firm growth. Although these regressions provide some statistical evidence that union elections help predict firm behavior, the lack of any union-win effect, coupled with the quite different effects for union losses across industries, renders suspect the conclusion that elections truly cause growth changes.

The evidence that firm growth causes election activity is even less compelling. The estimates of (2) in Table 6 provide scant evidence that changes in growth are followed by any measurable change in election activity. In fact, the only significant coefficients (at the 0.10 level) on a growth variable in all of Table 6 indicate that a shock to sales growth has a slight positive effect on successful election activity at manufacturing firms.[12] The remainder of the election VAR's give no indication that innovations in a firm's growth have any impact on union election activity at that firm. The F tests for causality imply that the hypothesis of noncausality from growth to elections cannot be rejected in any of the specifications.

The general implication of the VAR's is that there is only a weak relationship between union organizing activity and firm growth. In addition, the only rejections of noncausality imply that losing elections have opposite effects on growth in manufacturing and nonmanufacturing firms. Most importantly, there is no evidence that union success affects growth.

At this point it is instructive to consider why we find so little evidence of a causal link between union elections and firm growth. The most obvious candidate is the confounding effect of the business cycle. If firms react differently to cyclical shocks, then, even though we include controls for aggregate growth, the strong cyclical pattern of union elections makes it difficult to obtain precise estimates of the relationship between changes in unionization and changes in growth. In essence the econometrician is attempting to infer whether a firm's behavior would have been different over a certain range of the business cycle had it not experienced a union organizing drive. If there is sufficient heterogeneity in cyclical sensitivity across firms, more precise estimates of this effect can only be obtained by observing the same firm's behavior over a repeated number of business cycles. Given the opposite signs of the effects on growth and the similar cyclical election patterns across industries, however, this differing-sensitivity hypothesis requires that manufacturing firms be more sensitive and that nonmanufacturing firms be less sensitive to the

[11]The lone exception is a weak negative effect of winning elections on investment growth three years later in nonmanufacturing firms.

[12]The mean number of successful elections per firm/year is about 0.026, while the typical (median) fraction of employment involved in successful elections is 0.026. Multiplying these gives an estimated probability of 0.00068 that a typical worker is involved in a union success in a given year. The estimates in Table 6 imply that a 3-percentage-point increase in annual sales growth, which is approximately the mean sales growth, would increase this probability to 0.00074, or by about one-tenth. Thus, this lone significant effect is quite small in magnitude.

business cycle when they experience a union election.[13]

The fact that there are relatively few successful elections in the sample may also make it difficult to obtain precise estimates of union-win effects in either industry.[14] It is important to note, however, that Ruback and Zimmerman (1984) found much larger and stronger equity effects for union wins than for union losses in this sample. Moreover, these estimated equity effects did not differ much across industries (Ruback and Zimmerman (1984 pp. 1154–5).

Finally, it is possible that a firm may not respond much to the occurrence of a union election, whether a win or loss, because it has already adjusted its behavior given the probability of unionization that if faces. Recently, Abowd (1989b) and Bronars and Deere (1991) have provided some evidence that differences in firm behavior are indeed associated with differences in the likelihood of unionization.

III. Conclusions

This paper began by noting that if firms alter their behavior in response to union elections then the process of new unionization is inefficient. Such changes in behavior imply that the associated loss to shareholders exceeds the gain to the union. Growing firms are more attractive targets of unionization because of the greater rents and quasi-rents available to shareholders, but they may be less susceptible to unionization because of their ability to maintain higher wages and greater job security. Understanding of the causes and consequences of unionization can be enhanced by a careful econometric examination of the time-series relationship between union organizing activity and changes in input and output growth at the firm level.

Our analysis suggests that the likelihood of a union election is quite sensitive to aggregate economic activity, but essentially unrelated to a firm's own recent changes in output or inputs. Knowledge of past sales, employment, and investment changes across firms are of little help in predicting when or where union elections will occur. Our analysis also shows that after controlling for general time-period effects, union election activity has a weak effect on firm behavior. Sales growth does change somewhat following a union election, but these estimated effects are only apparent when unions are unsuccessful in their attempt to organize, and they are opposite in sign for manufacturing and nonmanufacturing firms.

These results lead us to the cautious conclusion that the observed changes in a firm's growth that follow a union election have little to do with unionization itself. We do note, however, that there is a significant, negative cross-section relationship between successful elections and firm growth in our sample. It is possible that this relationship is causal but that the confounding effects of the business cycle prevent us from uncovering any significant effects of elections. From this we conclude that cross-section analyses of unionization and firm behavior should be viewed with some skepticism.

Our results also suggest that the large equity losses uncovered by Ruback and Zimmerman (1984), for these same elections, are unrelated to systematic changes in firms' inputs or output. Ruback and Zimmerman (1984) found an average 4-percent equity loss during the month of a successful union election, and we find no evidence of effects on growth for these election events. We cannot reject the hypothesis that the

[13] Richard B. Freeman and Morris M. Kleiner (1990) compare firm behavior before and after union elections using a sample of elections from the 1980's and find modest declines in employment levels following a union election. They also include a set of control establishments and year dummy variables, but as with our results, it is not clear that this controls sufficiently for the influence of the business cycle.

[14] In addition, the fact that we limit our analysis to firms for which the full 20-year panel of data are available may lessen the chances of finding election effects. If business failures follow a union success, then our estimates understate the impact of elections. Lastly, there may be important intrafirm effects of unionization that our firm-level data cannot detect. For example, the unionization of one plant in a multiestablishment firm may result in a significant shift of production activity to nonunion plants with little overall effect on the firm. If such were the case, however, it is not clear why there would be large and permanent effects on the firm's equity value.

equity losses for these elections represent a simple transfer of wealth from shareholders to workers. This finding for changes in unionization is consistent with the conclusion of Abowd (1989a) and Clark (1984) that negotiations between an existing union and firm lead to efficient wealth transfers.

Appendix

The data used in the analysis were obtained from three sources. Richard Ruback and Martin Zimmerman kindly provided us with the data used in their 1984 paper. These data included information on the firm in which the election was held, the petition and certification months, the number of workers involved in the election, and the outcome of the election. We supplemented these data by extending the time period through 1984 and by including every election that occurred at all 157 firms in our sample during the period 1977–1984. The supplemental election data were derived from the election data tapes maintained by the National Labor Relations Board and, on occasion, from the published Monthly Election Reports of the NLRB. The data on sales, employment, and investment as well as industry attachment were taken directly from the COMPUSTAT data tapes. All data used in the paper are available from the authors upon request.

We give here a brief description of the 2SLS-GLS estimation procedure used for Tables 5 and 6. Further details and the programs used are available from the authors upon request. We take first differences of equations (1) and (2) to remove the firm-specific effects. Differencing induces serial correlation in the error terms, and it also creates a correlation between ΔY_{it-1} and $\Delta\varepsilon_{it}$ and between ΔX_{it-1} and $\Delta\nu_{it}$ due to the shared terms dated $t-1$. The serial correlation of the errors was addressed by using a generalized least-squares (GLS) transformation to obtain a scalar covariance matrix, and the correlation between the errors and regressors was addressed by utilizing an instrumental-variables procedure (on the transformed equations). As suggested by Holtz-Eakin et al. (1988), we utilize lags of Y and X as instruments for the differenced and transformed variables.

The GLS transformation required knowledge of the covariance matrices for $\Delta\varepsilon$ and $\Delta\nu$. Initial consistent coefficient estimates were obtained from applying two-stage least squares to the differenced versions of (1) and (2) where year dummies and lags 2 through $M+1$ of Y and X were used as the instrument set. The resulting residuals were then used to estimate the covariance matrices. The differenced variables were then transformed using the results of a Cholesky-type decomposition of the estimated covariance matrices to yield transformed equations with scalar covariance matrices. We assumed that the correlation of errors across firms is zero given the inclusion of time effects; hence, the covariance matrices are block-diagonal with identical blocks.

The last step in obtaining the reported estimates was to apply two-stage least squares to the differenced and then transformed versions of (1) and (2), where the instrument set again comprised year dummies plus lags 2 through $M+1$ of Y and X. The appropriate value of the lag length M was obtained by performing F tests on successive lag lengths, beginning with $M=4$. We selected $M=3$ as the final specification.

Lastly, Tables A1 and A2 report the OLS estimates of equations (1) and (2) for comparison with Tables 5 and 6 in the text and for greater ease of replication.

REFERENCES

Abowd, John M., (1989a) "The Effect of Wage Bargains on the Stock Market Value of the Firm," *American Economic Review*, September 1989, *79*, 774–800.

_____, (1989b) "The Effects of Differential Unionization Environments on the Pattern of Interindustry Investment," mimeo, Cornell University, 1989.

Allen, Steven, G., "Unionized Construction Workers are More Productive," *Quarterly Journal of Economics*, May 1984, *99*, 251–74.

TABLE A1—VECTOR AUTOREGRESSIONS (OLS): GROWTH

	Dependent variables					
	Growth in employment		Growth in investment		Growth in sales	
Explanatory variables	Coefficient	SE	Coefficient	SE	Coefficient	SE
Firms in Manufacturing:						
Lag-1 union loss	−1.61	1.17	−3.80	4.13	−2.15	1.12
Lag-2 union loss	−2.45	1.17	−0.66	4.12	−1.60	1.12
Lag-3 union loss	−1.26	1.17	−0.13	4.12	−1.90	1.12
Lag-1 union win	−3.78	5.65	12.90	19.89	−3.01	5.40
Lag-2 union win	−1.31	3.63	5.47	12.80	−1.63	3.47
Lag-3 union win	−1.17	3.48	0.08	12.27	1.08	3.33
Lag-1 employment	0.05	0.02	—	—	—	—
Lag-2 employment	−0.05	0.02	—	—	—	—
Lag-3 employment	−0.09	0.02	—	—	—	—
Lag-1 investment	—	—	−0.23	0.02	—	—
Lag-2 investment	—	—	−0.28	0.02	—	—
Lag-3 investment	—	—	−0.17	0.02	—	—
Lag-1 sales	—	—	—	—	0.13	0.02
Lag-2 sales	—	—	—	—	−0.09	0.02
Lag-3 sales	—	—	—	—	−0.10	0.02
F tests (*p* values in parentheses):						
All union-loss coefficients	2.10 (0.098)		0.30 (0.824)		2.40 (0.066)	
All union-win coefficients	0.21 (0.890)		0.19 (0.901)		0.22 (0.901)	
All election coefficients	1.17 (0.318)		0.25 (0.961)		1.33 (0.242)	
Firms in Nonmanufacturing:						
Lag-1 union loss	1.27	2.18	8.18	9.56	6.62	2.61
Lag-2 union loss	1.54	2.16	15.55	9.54	7.88	2.61
Lag-3 union loss	1.76	2.11	10.86	9.32	4.68	2.56
Lag-1 union win	2.72	4.91	20.74	21.70	−3.15	5.91
Lag-2 union win	3.31	4.89	17.04	21.71	1.59	5.90
Lag-3 union win	−1.26	1.20	−10.01	5.56	0.34	1.46
Lag-1 employment	0.07	0.04	—	—	—	—
Lag-2 employment	0.03	0.04	—	—	—	—
Lag-3 employment	0.01	0.04	—	—	—	—
Lag-1 investment	—	—	−0.33	0.05	—	—
Lag-2 investment	—	—	−0.17	0.05	—	—
Lag-3 investment	—	—	−0.06	0.05	—	—
Lag-1 sales	—	—	—	—	0.11	0.05
Lag-2 sales	—	—	—	—	−0.11	0.04
Lag-3 sales	—	—	—	—	0.07	0.05
F tests (*p* values in parentheses):						
All union loss coefficients	0.46 (0.710)		1.42 (0.237)		5.57 (0.001)	
All union win coefficients	0.63 (0.594)		1.61 (0.186)		0.15 (0.931)	
All election coefficients	0.55 (0.769)		1.53 (0.168)		2.86 (0.010)	

Notes: Each regression also included a complete set of year and firm dummy variables. The coefficients on the union variables have been multiplied by 10. For firms in manufacturing, $N = 1{,}968$; for firms in nonmanufacturing, $N = 544$.

TABLE A2—VECTOR AUTOREGRESSIONS (OLS): ELECTIONS
(STANDARD ERRORS IN PARENTHESES)

	Dependent variables					
	Elections lost by the union			Elections won by the union		
Explanatory variables	(i)	(ii)	(iii)	(i)	(ii)	(iii)
Firms in Manufacturing:						
Lag-1 union loss	−11.64	−11.59	−11.63	0.51	0.54	0.52
	(2.32)	(2.32)	(2.32)	(0.49)	(0.49)	(0.49)
Lag-2 union loss	−10.73	−10.66	−10.76	0.99	1.04	0.97
	(2.32)	(2.32)	(2.32)	(0.49)	(0.49)	(0.49)
Lag-3 union loss	−11.75	−11.88	−11.61	0.18	0.04	0.20
	(2.32)	(2.32)	(2.32)	(0.49)	(0.49)	(0.49)
Lag-1 union win	−5.39	−5.00	−5.00	−7.92	−7.88	−7.88
	(11.20)	(11.20)	(11.20)	(2.38)	(2.38)	(2.38)
Lag-2 union win	1.81	1.92	1.71	−3.68	−3.56	−3.69
	(7.21)	(7.21)	(7.20)	(1.53)	(1.53)	(1.53)
Lag-3 union win	−1.96	−2.28	−1.81	−4.00	−4.06	−3.99
	(6.91)	(6.91)	(6.91)	(1.47)	(1.47)	(1.47)
Lag-1 employment	−0.61	—	—	0.24	—	—
	(4.57)			(0.97)		
Lag-2 employment	−1.27	—	—	0.63	—	—
	(4.44)			(0.94)		
Lag-3 employment	0.24	—	—	−0.74	—	—
	(4.39)			(0.93)		
Lag-1 investment	—	0.38	—	—	0.33	—
		(1.29)			(0.27)	
Lag-2 investment	—	−0.03	—	—	0.26	—
		(1.29)			(0.27)	
Lag-3 investment	—	−1.70	—	—	−0.15	—
		(1.32)			(0.28)	
Lag-1 sales	—	—	1.08	—	—	−0.42
			(4.84)			(1.03)
Lag-2 sales	—	—	5.02	—	—	1.69
			(4.60)			(0.98)
Lag-3 sales	—	—	−4.61	—	—	−1.26
			(4.58)			(0.97)
F tests (*p* values in parentheses):						
All employment coefficients	0.04	—	—	0.36	—	—
	(0.991)			(0.785)		
All investment coefficients	—	0.68	—	—	0.98	—
		(0.562)			(0.402)	
All sales coefficients	—	—	0.71	—	—	1.37
			(0.547)			(0.251)

TABLE A2—*Continued*

	Dependent variables					
	Elections lost by the union			Elections won by the union		
Explanatory variables	(i)	(ii)	(iii)	(i)	(ii)	(iii)
Firms in Nonmanufacturing:						
Lag-1 union loss	−6.94	−6.54	−6.73	−0.04	0.01	0.06
	(4.57)	(4.54)	(4.55)	(0.65)	(0.64)	(0.64)
Lag-2 union loss	−5.29	−5.99	−5.55	0.11	0.04	0.07
	(4.53)	(4.53)	(4.55)	(0.64)	(0.64)	(0.64)
Lag-3 union loss	−4.29	−4.08	−4.23	−0.42	−0.41	−0.34
	(4.43)	(4.43)	(4.46)	(0.63)	(0.63)	(0.63)
Lag-1 union win	−1.32	−1.16	−1.35	−0.92	−0.96	−0.94
	(10.30)	(10.31)	(10.30)	(1.46)	(1.46)	(1.46)
Lag-2 union win	3.75	4.22	3.46	−1.21	−1.26	−1.19
	(10.27)	(10.31)	(10.28)	(1.45)	(1.46)	(1.46)
Lag-3 union win	0.40	0.53	0.19	−0.04	−0.13	−0.07
	(2.53)	(2.64)	(2.55)	(0.36)	(0.37)	(0.36)
Lag-1 employment	−1.28	—	—	1.72	—	—
	(9.02)			(1.28)		
Lag-2 employment	−13.84	—	—	−1.21	—	—
	(8.40)			(1.19)		
Lag-3 employment	−0.89	—	—	0.46	—	—
	(8.31)			(1.18)		
Lag-1 investment	—	−1.57	—	—	0.40	—
		(2.21)			(0.31)	
Lag-2 investment	—	−0.36	—	—	0.13	—
		(2.28)			(0.32)	
Lag-3 investment	—	−1.39	—	—	0.04	—
		(2.17)			(0.31)	
Lag-1 sales	—	—	−0.65	—	—	0.01
			(7.95)			(1.13)
Lag-2 sales	—	—	−8.55	—	—	−0.16
			(7.82)			(1.11)
Lag-3 sales	—	—	4.84	—	—	−0.23
			(8.01)			(1.14)
F tests (*p* values in parentheses):						
All employment coefficients	0.92	—	—	1.01	—	—
	(0.432)			(0.388)		
All investment coefficients	—	0.29	—	—	0.56	—
		(0.836)			(0.643)	
All sales coefficients	—	—	0.50	—	—	0.02
			(0.686)			(0.995)

Notes: Each regression also included a complete set of year and firm dummy variables. The coefficients on the election and growth variables have been multiplied by 100 and 1,000, respectively. For firms in manufacturing, $N = 1{,}968$; for firms in nonmanufacturing, $N = 544$.

_____, "Can Union Labor Ever Cost Less?" *Quarterly Journal of Economics*, May 1987, *102*, 347–73.

Ashenfelter, Orley C. and Pencavel, John H., "American Trade Union Growth: 1900–1960," *Quarterly Journal of Economics*, August 1969, *83*, 434–48.

Blanchflower, David G., Millward, Neil and Oswald, Andrew J., "Unionism and Employment Behavior," *Economic Journal*, July 1991, *101*, 815–34.

Bronars, Stephen G. and Deere, Donald R., "Union Representation Elections and Firm Profitability," *Industrial Relations*, Winter 1990, *29*, 15–37.

_____ **and** _____, "The Threat of Unionization, the Use of Debt, and the Preservation of Shareholder Wealth," *Quarterly Journal of Economics*, February 1991, *106*, 231–54.

Brown, Charles and Medoff, James L., "Trade Unions in the Production Process," *Journal of Political Economy*, June 1978, *86*, 355–78.

Brown, James N. and Ashenfelter, Orley C., "Testing the Efficiency of Employment Contracts," *Journal of Political Economy*, June 1986, *93*, S40–S87.

Card, David, "Efficient Contracts with Costly Adjustment: Short-Run Employment Determination for Airline Mechanics," *American Economic Review*, December 1986, *76*, 1045–71.

Clark, Kim B., "The Impact of Unionization on Productivity: A Case Study," *Industrial and Labor Relations Review*, July 1980, *33*, 451–69.

_____, "Unionization and Firm Performance: The Impact on Profits, Growth, and Productivity," *American Economic Review*, December 1984, *74*, 893–919.

Connolly, Robert A., Hirsch, Barry T. and Hirschey, Mark, "Union Rent Seeking, Intangible Capital, and the Market Value of the Firm," *Review of Economics and Statistics*, November 1986, *68*, 567–77.

Eberts, Randall W. and Stone, Joe A., "On the Contract Curve: A Test of Alternative Models of Collective Bargaining," *Journal of Labor Economics*, January 1986, *4*, 66–81.

Freeman, Richard and Kleiner, Morris M., "The Impact of New Unionization on Wages and Working Conditions," *Journal of Labor Economics*, January 1990, (Part 2), *8*, S8–S25.

Hirsch, Barry T., "Union Coverage and Profitability Among U.S. Firms," *Review of Economics and Statistics*, May 1991, *73*, 69–77.

Holtz-Eakin, Douglas, Newey, Whitney K. and Rosen, Harvey S, "Estimating Vector Autoregressions with Panel Data," *Econometrica*, November 1988, *56*, 1371–95.

Karier, Thomas, "Unions and Monopoly Profits," *Review of Economics and Statistics*, February 1985, *67*, 34–42.

Leonard, Jonathan S., "Unions and Employment Growth," *Industrial Relations*, Winter 1992, *31*, 80–94.

Machin, Stephen J. and Wadhwani, Sushil B., "The Effects of Unions on Organisational Change and Employment," *Economic Journal*, July 1991, *101*, 835–54.

MaCurdy, Thomas E. and Pencavel, John H., "Testing Between Competing Models of Wage and Employment Determination in Unionized Markets," *Journal of Political Economy*, June 1986, *93*, S3–S39.

Montgomery, Edward B., "Employment and Unemployment Effects of Unions," *Journal of Labor Economics*, April 1989, *7*, 170–90.

Nickell, Stephen J., "Biases in Dynamic Models with Fixed Effects," *Econometrica*, September 1981, *49*, 1417–26.

Rose, Nancy L., "Labor Rent Sharing and Regulation: Evidence from the Trucking Industry," *Journal of Political Economy*, December 1987, *95*, 1146–78.

Ruback, Richard S. and Zimmerman, Martin B., "Unionization and Profitability: Evidence from the Capital Market," *Journal of Political Economy*, December 1984, *92*, 1134–57.

Salinger, Michael A., "Tobin's *q*, Unionization, and the Concentration–Profits Relationship," *Rand Journal of Economics*, Summer 1984, *15*, 159–70.

Voos, Paula B. and Mishel, Lawrence R., "The Union Impact on Profits: Evidence from Industry Price–Cost Margin Data," *Journal of Labor Economics*, January 1986, *4*, 105–33.

Seniority and Monopsony in the Academic Labor Market

By MICHAEL R RANSOM*

Workers with high seniority usually have higher wages than workers with low seniority and the same total labor-market experience. In contrast, the results of this paper indicate that for university professors higher seniority is associated with lower salaries. I document this finding for national surveys and individual institutions and explain it as due to monopsonistic discrimination by employers. (*JEL* J31, J44)

An idea that is central to many theories about the life-cycle structure of pay is that workers with high seniority are paid more than otherwise similar workers with low seniority. As an example of the types of returns to seniority observed in cross-sectional studies, Table 1 reports results of a regression analysis of various occupational and educational groups from a recent sample of adult workers in the United States, the May 1988 Current Population Survey (CPS).[1] For all workers, earnings initially increase by about 3 percent with each additional year of seniority. For a worker with 10 years of seniority (sample average seniority is 9.8 years), an additional year of seniority is associated with a 2-percent increase in pay, and pay is more than 30-percent higher than for a worker of the same age but with no seniority. Wages show a strong direct association with seniority.

For most of the other groups listed in Table 1, the seniority profile is quite similar. Even for highly educated workers and for workers who have very "general" skills, such as accountants and school teachers, a substantial part of salary growth is associated with (if not attributable to) increases in seniority. For school teachers, seniority appears to be particularly important in explaining the life cycle of pay, while experience seems relatively unimportant.

The labor market for university professors provides a curious contrast. In the small sample of college and university teachers presented in Table 1, there is little evidence that seniority matters at all. However, careful analysis of large samples of university professors shows that salaries and seniority are *negatively* correlated, after controlling for experience and other variables. Anecdotal evidence suggests that this can be extreme at times. In a recent survey by the American Council on Education (see Debra E. Blum, 1990), more than half of the doctoral-granting institutions surveyed responded that in some departments they had hired "new, junior faculty members at a salary above that of some senior faculty members in the same department." In another recent example, two professors filed suit under the age discrimination act, alleging that the practice of paying "market rates" to new hires but not to others had a disparate impact on older workers. The two plaintiffs were the most senior members of the school's business college and were the *lowest* paid (see Bureau of National Affairs, 1991).[2]

*Department of Economics, Brigham Young University, Provo, UT 84602. I am grateful for helpful discussions with David Card, Val Lambson, James B. McDonald, John Pencavel, and participants of the Industrial Relations seminar at Princeton University. Some of the data used in this paper were provided by the Interuniversity Consortium for Political and Social Research. I am solely responsible for opinions and analyses contained in the paper.

[1]Technical documentation for this survey can be found in U.S. Department of Commerce (1989).

[2]The employees lost the case.

TABLE 1—SELECTED REGRESSION COEFFICIENTS, MAY 1988 CURRENT POPULATION SURVEY
(DEPENDENT VARIABLE = LOGARITHM OF WEEKLY EARNINGS)

Group	Age	$(Age)^2$	Seniority	$(Seniority)^2$	Marginal effect of seniority at 10 years	Sample size
All workers	0.0395	-4.66×10^{-4}	0.0315	-5.23×10^{-4}	0.0202	16,404
	(0.0027)	(3.10×10^{-5})	(0.0013)	(4.20×10^{-5})		
Highly educated (18 or more years of schooling)	0.0678	-7.16×10^{-4}	0.0224	-3.56×10^{-4}	0.0153	1,457
	(0.0119)	(1.36×10^{-4})	(0.0056)	(2.08×10^{-4})		
School teachers [156–157]	−0.0041	2.36×10^{-5}	0.0366	-6.14×10^{-4}	0.0243	572
	(0.0146)	(1.71×10^{-4})	(0.0068)	(2.38×10^{-4})		
Accountants [23]	0.0936	−0.0011	0.0251	3.69×10^{-4}	0.0176	204
	(0.0240)	(0.0003)	(0.0117)	(4.08×10^{-4})		
College and university teachers [113–154]	0.1381	−0.0013	−0.0030	4.78×10^{-4}	0.0066	115
	(0.0431)	(0.0005)	(0.0211)	(7.72×10^{-4})		

Notes: Standard errors are in parentheses. Numbers in brackets refer to the CPS occupational codes for the group. The sample consists of workers between the ages of 25 and 70 who responded to job-seniority questions. Weekly earnings data are right-censored (at $999) so a "Tobit" regression is used. Weekly earnings for about half of the sample come from the June 1988 Current Population Survey. Other variables included in the analysis are education level, race, sex, whether Hispanic, whether resident of a metropolitan area, and region of the country.

I. The Seniority Profile of Compensation

There are many theories of the labor market that predict a positive association between pay and seniority. Perhaps the most prominent explanation is the theory of specific human capital, explained in Walter Oi (1962), Jacob Mincer (1974), and many others. According to this theory, the firm rewards the worker for acquiring skills and knowledge that are specific to the current employer. Higher wages deter the worker from leaving with the firm's portion of the human-capital investment.

More recently, Gary S. Becker and George Stigler (1974) and Edward P. Lazear (1981) have developed models in which the firm pays higher wages to more senior workers as an incentive mechanism. The promise of high pay motivates the worker to provide an appropriate level of effort, since workers who are caught shirking or cheating are dismissed from the firm and lose the reward of higher wages later in their careers. Unlike the human-capital model, this model does not predict higher productivity for more senior workers.

In the model of Joanne Salop and Steven Salop (1976), firms offer steeply sloped seniority profiles as a selection mechanism. The present value of the earnings stream is higher than alternatives, but the starting pay is lower than alternatives. The high wages in late career compensate for the low starting wages only for employees who intend to stay with the firm for a long period of time.

Numerous empirical studies of wages, as in the analysis reported in Table 1, have found a positively sloped seniority profile. However, a positive correlation between pay and seniority might be observed in a cross section even if pay does not increase with seniority for individuals. For example, Boyan Jovanovic (1979) suggests a matching model in which firms and workers learn about each other over time. Well-matched workers are highly paid and stay with the firm. Poorly matched workers receive low wages and are likely to leave the firm searching for better

matches. As a result, *average* wages increase with seniority even if the wages of any individual do not. Efficiency wage models predict a similar pattern.[3] Because individuals are paid above alternatives, they are less likely to quit, so workers in high-wage firms will have higher seniority on average than workers in low-wage firms. Indeed, for some versions of the efficiency wage model, lower costs associated with lower turnover are crucial.

Some recent empirical work has tried to determine whether wages increase with seniority by studying longitudinal changes in the pay of individuals. In similar papers, Katherine G. Abraham and Henry S. Farber (1987) and Joseph G. Altonji and Robert A. Shakotko (1987) concluded that much of the cross-sectional correlation between wages and seniority is due to heterogeneity, a conclusion that is consistent with efficiency wage models or other types of matching models. However, Robert Topel (1991) has criticized these two papers. Using the same longitudinal data set but slightly different estimation methods, Topel finds that wages of individuals increase rather sharply with increases in seniority. Topel reports that for a typical male worker in the United States, 10 years of seniority increase the wage by more than 25 percent relative to alternatives. This is only slightly smaller than the cross-sectional estimate reported in Table 1 for all workers.

II. The Seniority Profile for University Professors

Findings from the labor market for university professors are inconsistent with all of these theories and estimates. Here I present evidence from several sources that demonstrates a negative return to seniority for university professors. Data are from two different types of sources: first, from nationwide surveys of university faculty; and second, from employment records of individual institutions.

[3]Joseph Stiglitz (1987) describes several versions of the efficiency wage model.

TABLE 2—SELECTED REGRESSION COEFFICIENTS FROM 1969 CARNEGIE SURVEY DATA (DEPENDENT VARIABLE = LOGARITHM OF ANNUAL SALARY)

Variable	All universities and colleges	"High-quality" universities only
Experience:		
0–1 years	—	—
2–3 years	0.099**	0.170**
	(0.006)	(0.015)
4–6 years	0.186**	0.280**
	(0.006)	(0.015)
7–9 years	0.275**	0.387**
	(0.007)	(0.016)
10–14 years	0.383**	0.528**
	(0.007)	(0.016)
15–19 years	0.477**	0.662**
	(0.007)	(0.017)
20–29 years	0.552**	0.760**
	(0.008)	(0.018)
30 or more years	0.591**	0.851**
	(0.010)	(0.027)
Seniority:		
0–1 years	—	—
2–3 years	−0.026**	−0.030**
	(0.005)	(0.012)
4–6 years	−0.030**	−0.040**
	(0.005)	(0.012)
7–9 years	−0.026**	−0.026
	(0.006)	(0.014)
10–14 years	−0.049**	−0.056**
	(0.006)	(0.015)
15–19 years	−0.073**	−0.110**
	(0.007)	(0.017)
20–29 years	−0.067**	−0.128**
	(0.008)	(0.018)
30 or more years	−0.042**	−0.154**
	(0.011)	(0.026)
Sample size:	50,214	11,665
R^2:	0.544	0.504

Notes: Standard errors are given in parentheses. Other variables: degree type, field of teaching appointment, sex, race, quality of degree-granting institution, private or public employer, and length of contract.

**Coefficient is significantly different from 0 at the 1-percent level.

A. *Estimates from National Surveys*

Tables 2, 3, and 4 report estimates from a salary regression model on data from three national surveys of teaching faculty at universities and colleges. I report results for a sample of all college and university faculty and for a smaller sample of faculty at research universities. Table 2 refers to data

TABLE 3—SELECTED REGRESSION COEFFICIENTS, 1972–1973 ACE SURVEY DATA (DEPENDENT VARIABLE = LOGARITHM OF ANNUAL SALARY)

Variable	All universities and colleges	Research universities only
Experience	0.039**	0.045**
	(0.001)	(0.001)
(Experience)^2	-6.3×10^{-4}*	-7.0×10^{-4}*
	(1.8×10^{-5})	(2.1×10^{-5})
Seniority	-3.5×10^{-4}	-0.0037**
	(7.3×10^{-4})	(0.0009)
(Seniority)^2	9.2×10^{-6}**	5.4×10^{-5}*
	(2.2×10^{-6})	(2.6×10^{-5})
Marginal effect of an additional year:		
At 5 years of seniority	-3.0×10^{-4}	-0.0034**
	(6.4×10^{-4})	(0.0008)
At 10 years of seniority	-2.6×10^{-4}	-0.0031**
	(5.4×10^{-4})	(0.0007)
Sample size:	30,313	16,207
R^2:	0.464	0.540

Notes: Standard errors are given in parentheses. Other variables: degree held, race, sex, field, region, city size, SMSA type, and contract length.

*Coefficient is significantly different from 0 at the 5-percent significance level.

**Coefficient is significantly different from 0 at the 1-percent significance level.

TABLE 4—SELECTED REGRESSION COEFFICIENTS, 1977 SURVEY OF THE AMERICAN PROFESSORATE DATA (DEPENDENT VARIABLE = LOGARITHM OF ANNUAL SALARY)

Variable	All universities and colleges	Research universities only
Experience:		
0–2 years	—	—
3–5 years	0.084**	0.099**
	(0.021)	(0.029)
6–10 years	0.196**	0.249**
	(0.021)	(0.029)
11–15 years	0.351**	0.405**
	(0.023)	(0.031)
16–20 years	0.476**	0.549**
	(0.024)	(0.033)
More than 20 years	0.545**	0.610**
	(0.025)	(0.034)
Seniority:		
0–2 years	—	—
3–5 years	0.014	0.006
	(0.016)	(0.022)
6–10 years	0.026	−0.014
	(0.017)	(0.023)
11–15 years	−0.001	−0.042
	(0.019)	(0.026)
16–20 years	−0.013	−0.069*
	(0.022)	(0.029)
More than 20 years	0.010	−0.040
	(0.023)	(0.030)
Sample size:	3,477	2,232
R^2:	0.574	0.567

Notes: Standard errors are given in parentheses. Other variables: degree type, field of teaching appointment, race, sex, region, quality of degree-granting institution, and contract length (8–9 months or 11–12 months).

*Coefficient is significantly different from 0 at the 5-percent significance level.

**Coefficient is significantly different from 0 at the 1-percent significance level.

from the Carnegie Commission National Survey of Higher Education: Faculty Study, 1969.[4] This survey was completed in the spring of 1969. There were a total of 60,028 respondents, of which 50,214 provided complete responses to questions about salary, experience, seniority, and other factors considered in the regression model. The survey reports experience and seniority in intervals only.[5]

Experience shows a strong positive association with salary. For all colleges and universities, salaries increase by about 4 or 5 percent per year initially, with diminishing marginal returns. Experience profiles are much steeper at "high-quality universities."[6]

[4]This data set is available from the Interuniversity Consortium for Political and Social Research and is documented in Carnegie Commission (1978).

[5]Experience is here defined as the number of years employed in colleges and universities beyond the level of a teaching or research assistant. Seniority is defined as the number of years employed at current institution beyond the level of a teaching or research assistant.

[6]Since the 1969 study by the Carnegie Commission, universities and colleges in the United States have been classified into the following categories: Research Universities I and II, Doctoral-Granting Universities I and II, Comprehensive Colleges and Universities I and II, Liberal-Arts Colleges I and II, and Two-Year Colleges. These categories are determined by the number and type of degrees granted and the amount of research funding received by the universities. For the 1969 survey, I have assumed that the "high-quality" university category corresponds roughly to the more recent designation of Research University.

On the other hand, the marginal effect of seniority is clearly negative. For the overall sample, individuals with 10–14 years of seniority earn about 7 percent less than individuals with less than 2 years of seniority, after controlling for type of degree, field, sex, race, quality of institution that granted the individual his or her highest degree, the length of contract (8–9 months versus 11–12 months), and whether the employing institution is privately controlled. At high-quality universities the effect is even more pronounced. Salaries decline by about 0.5 percent per year of seniority. Individuals with at least 30 years of seniority earned approximately 15-percent less than individuals with less than 2 years of seniority.

The results reported in Table 3 are based on the American Council on Education's (ACE) 1973 survey, which was conducted between December 1972 and June 1973.[7] This survey was designed to parallel closely the 1969 Carnegie Commission survey. There were 53,029 respondents, of which 30,313 provided complete information on the variables used in the regression models. In this survey, experience is defined as years since the highest degree was granted. Seniority is measured as the number of years since an individual was hired by the current institution.

Table 3 reports regression estimates of the effect of seniority and experience for the 1973 data. Since the variables are measured in continuous form, I have estimated a model with linear and quadratic measures of both variables. Salaries increase rapidly with experience, but with diminishing marginal returns. The rate of growth with experience is larger for individuals at research universities. The estimated effect of seniority is negative, again. However, for the overall sample the hypothesis that seniority has no effect cannot be rejected at usual significance levels.[8] For the sample from research universities, salaries decline by about 0.4 percent per year with each additional year of seniority. The decline is roughly linear, as in the 1969 sample. An additional year of seniority reduces the salary of someone with 10 years of seniority by about 0.3 percent.

Table 4 refers to data from the "1977 Survey of the American Professoriate," administered by the Institute for Social Inquiry at the University of Connecticut. This survey had only 4,248 respondents, of which 3,477 provided complete responses for the variables used in this analysis. Data on experience and seniority are available only in interval form.[9]

The 1977 data display patterns similar to the other data sets. Salaries grow with experience, and this growth is more rapid at research universities. The data for all universities show essentially no correlation between seniority and salary after controlling for experience and other variables. For research universities, salaries do fall with increased seniority. However, the effect is smaller than that observed in the 1969 and 1973 data. The effect is largest for those with 16–20 years seniority, who had salaries that were about 7-percent lower than those with 2 years or less seniority, after controlling for experience, education, and other factors.

Figure 1 graphically summarizes the effect of seniority on salary at research universities from the analysis presented in the last columns of Tables 2, 3, and 4. The 1973 and 1977 surveys suggest that the most a professor could gain by changing employers is about 5 or 6 percent. The 1969 survey shows a much larger potential seniority "penalty."

A possible weakness of the preceding analysis is that academic labor markets are field-specific. The demand and supply of

[7]See Alan E. Bayer (1973) for a detailed discussion of the survey. This documentation and the data set are available through the Interuniversity Consortium for Political and Social Research.

[8]If seniority is redefined in the same intervals used in Table 2, the lowest-seniority group is the highest-paid group. Salaries for each of the other seniority intervals are lower than for the lowest-seniority interval, and the hypothesis that there is no difference between the lowest-seniority group and the higher-seniority groups can be rejected with usual tests. However, the seniority "penalty" is about the same for all seniority levels.

[9]For this data, experience is defined as years employed on a full-time basis in higher education. Seniority is defined as years employed on a full-time basis at the present institution.

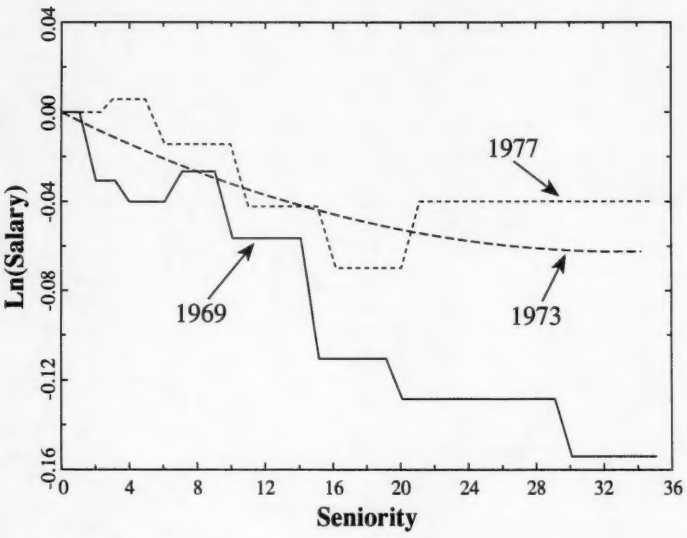


FIGURE 1. ESTIMATED SENIORITY–SALARY PROFILE FOR FACULTY OF RESEARCH UNIVERSITIES: NATIONAL SURVEY DATA, VARIOUS YEARS

accounting professors may be quite different from that of history professors. To analyze salary determination in different academic fields, Table 5 presents estimates for selected narrowly defined fields based on the 1973 American Council of Education survey, restricting the sample to faculty at research universities.[10] At 10 years of seniority, the effect of an additional year of seniority is negative for all fields except elementary and secondary education and chemistry. Most estimates in Table 5 indicate a decline of more than 0.5 percent per year, and some, such as accounting, and economics show a decline of more than 1 percent per year.[11] (Average seniority ranges from approximately 9 years to 12 years for these groups.)

[10]These models are estimated for 15 of the 75 fields that could be designated in the survey. Fields were selected to represent a broad range of specialities, but also to obtain a large sample size. Field refers to the department of the teaching appointment of the respondent.

[11]Since the relationship estimated is quadratic, the magnitude of the seniority effect usually falls with higher seniority. However, in most cases the quadratic term is quite small, so the relationship is almost linear.

B. *Estimates from Individual Institutions*

Estimates from individual universities also show negatively sloped seniority profiles. For example, Emily Hoffman (1976) estimated a salary regression model with a linear seniority term for data from two large universities. She found a decline of 0.2 percent per year at the University of Massachusetts using 1974 data, and a decline of 0.5 percent per year at another "large urban university," using data from 1970. Her analysis controlled for age (not experience), sex, race, field, and degree type. (The University of Massachusetts is categorized as a "research university" under the Carnegie system. Presumably the other "large urban university" is comparable.)

I have estimated similar models for data from another large state university, the University of Arizona.[12] Table 6 presents estimates for salaries from the 1972, 1977, and

[12]See Sharon B. Megdal and Ransom (1985) for a description of the data on which Table 6 is based. Copies of the data set are available from the author upon request. The University of Arizona is categorized a "Research University" under the Carnegie system.

TABLE 5—SELECTED REGRESSION COEFFICIENTS FROM 1973 ACE SURVEY DATA: ESTIMATES BY DEPARTMENT OF TEACHING APPOINTMENT (DEPENDENT VARIABLE = LOGARITHM OF ANNUAL SALARY)

Field	Experience	(Experience)^2	Seniority	(Seniority)^2	Marginal effect of seniority at 10 years	R^2	Sample size
Microbiology	0.0620**	−0.0011**	−0.0138*	0.0003	−0.0122*	0.570	237
	(0.0080)	(0.0002)	(0.0070)	(0.0002)	(0.0062)		
History	0.0542**	−0.0007**	−0.0070	-2.4×10^{-5}	−0.0073*	0.668	468
	(0.0038)	(0.0001)	(0.0046)	(1.4×10^{-4})	(0.0034)		
Mathematics and statistics	0.0684**	−0.0011**	−0.0087*	0.0001	−0.0077*	0.692	717
	(0.0035)	(0.0001)	(0.0036)	(0.0001)	(0.0027)		
Chemistry	0.0650**	−0.0010**	0.0052	−0.0003	0.0024	0.638	494
	(0.0058)	(0.0001)	(0.0057)	(0.0002)	(0.0042)		
General psychology	0.0602**	−0.0009**	−0.0078	-4.0×10^{-6}	−0.0079	0.722	328
	(0.0046)	(0.0001)	(0.0056)	(0.0002)	(0.0041)		
Economics	0.0560**	−0.0009**	−0.0094	-3.1×10^{-5}	−0.0097**	0.635	379
	(0.0041)	(0.0001)	(0.0050)	(1.0×10^{-4})	(0.0037)		
Sociology	0.0510**	−0.0008**	0.0004	−0.0002	−0.0019	0.650	287
	(0.0052)	(0.0002)	(0.0059)	(0.0002)	(0.0042)		
Accounting	0.0604**	−0.0013**	−0.0263**	0.0007*	−0.0189**	0.725	110
	(0.0084)	(0.0002)	(0.0086)	(0.0002)	(0.0065)		
Management	0.0427**	−0.0007**	−0.0086	−0.0002	−0.0067	0.440	153
	(0.0100)	(0.0002)	(0.0105)	(0.0003)	(0.0081)		
Elementary and secondary education	0.0267**	−0.0003**	0.0036	−0.0001	0.0023	0.667	233
	(0.0054)	(0.0001)	(0.0061)	(0.0002)	(0.0045)		
Chemical engineering	0.0621**	−0.0011**	−0.0121*	0.0003	−0.0090*	0.753	168
	(0.0097)	(0.0001)	(0.0058)	(0.0002)	(0.0043)		
Music	0.0260**	−0.0003	−0.0085	0.0003	−0.0054	0.434	269
	(0.0064)	(0.0002)	(0.0063)	(0.0002)	(0.0049)		
Medicine or surgery	0.0534**	−0.0009**	−0.0023	−0.0001	−0.0037	0.261	750
	(0.0073)	(0.0002)	(0.0072)	(0.0002)	(0.0051)		
Nursing	0.0180**	−0.0002	−0.0123*	0.0006**	−0.0060	0.691	210
	(0.0067)	(0.0062)	(0.0059)	(0.0002)	(0.0044)		
English language and literature	0.0519**	−0.0008**	−0.0114*	0.0002	−0.0093*	0.521	670
	(0.0046)	(0.0001)	(0.0053)	(0.0002)	(0.0039)		

Note: Standard errors are given in parentheses.
*Coefficient is significantly different from 0 at the 5-percent significance level.
**Coefficient is significantly different from 0 at the 1-percent significance level.

1982 academic years. The estimated coefficients indicate that, initially, an additional year of seniority reduces salary by 0.5–1.5 percent. The results from 1982 suggest that salaries decline at a rate of about 1 percent per year, a much larger effect than observed in the national samples. The estimated seniority profile "bottoms out" at about 15 years for the 1972 sample and at about 25 years for the 1977 sample. Figure 2 graphically shows the corresponding estimated seniority–salary profiles.

In summary, data from large research-oriented universities in the United Sates show that individuals with high seniority are paid less, on average, than individuals with low seniority, after controlling for total experience, sex, field, and other variables.[13]

[13]For other types of institutions, the effect of seniority on salaries is not so clear. For example, analysis of separate samples estimates positive returns to seniority for junior-college faculty in the 1969 and 1973 data, negative returns for liberal-arts colleges in the 1977

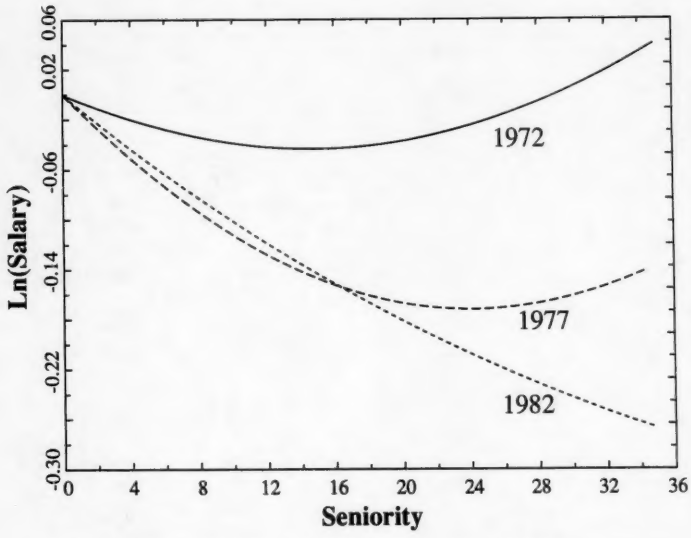


FIGURE 2. ESTIMATED SENIORITY–SALARY PROFILES FOR FACULTY OF THE UNIVERSITY OF ARIZONA, VARIOUS YEARS

TABLE 6—SELECTED REGRESSION COEFFICIENTS FROM ANALYSIS OF UNIVERSITY OF ARIZONA SALARY DATA (DEPENDENT VARIABLE = LOGARITHM OF ANNUAL SALARY)

Variable	1972	1977	1982
Years experience	0.0317**	0.0440**	0.0439**
	(0.0018)	(0.0017)	(0.0021)
(Experience)^2	-0.0005**	-0.0006**	-0.0006**
	(4.0×10^{-5})	(3.9×10^{-5})	(5.3×10^{-5})
Years seniority	-0.0058**	-0.0144**	-0.0111**
	(0.0020)	(0.0019)	(0.0022)
(Seniority)^2	0.0002*	0.0003**	0.0001
	(7.4×10^{-5})	(6.3×10^{-5})	(5.6×10^{-5})
Other variables	whether Ph.D., sex, 52 department indicators	whether Ph.D., sex, 44 department indicators	whether Ph.D., sex, 87 department indicators
Sample size:	1,085	1,127	1,197
R^2:	0.5895	0.6737	0.6407

Note: Standard errors are given in parentheses.
*Coefficient is significantly different from 0 at the 5-percent significance level.
**Coefficient is significantly different from 0 at the 1-percent significance level.

While estimates vary from sample to sample and from year to year, it appears that salaries decline by about 0.5 percent per year of seniority. For a typical professor, with about 10 years of seniority, changing employers would result in an increase in pay of from around 5 percent (using estimates from the national samples) to about 10 per-

data, and positive returns for comprehensive colleges and universities in the 1977 data.

cent (using estimates from the University of Arizona sample). This analysis only considers salaries. Nonmoney compensation at universities is also probably much less tied to seniority than at the typical firm. For example, one of the biggest components of in-kind pay is paid leave. In the United States, paid leave is traditionally tied to seniority. In most academic institutions, leave time is similar for professors of all seniority levels.

Durable employment relationships at universities are associated with low salaries! This result is astonishing in the face of the current economic theories of the seniority–pay relation. The human-capital model, for example, suggests no reason why pay would fall with higher seniority, even in cases where there is no "specific" human capital. Likewise, the well-known matching and deferred-compensation models predict that durable employment relationships will be high-paying.

III. Worker Quality and Seniority

One possible explanation for the negative effect of seniority on salaries in cross sections is that individuals with long seniority are simply less productive. For example, one argument is that the best professors can find employment anywhere, but the less able have to stay where they are. This is the basic flavor of two models that have appeared in the literature. Milton Harris and Bengt Holmstrom (1982) suggest a model in which a worker of unknown ability is initially offered a guaranteed lifetime wage. Over time, the worker's productivity is revealed to all firms. If a worker turns out to be very productive, the wage will be increased to match outside offers. However, if an individual turns out to have low productivity, no outside offer will dominate the initial guaranteed wage. Those receiving outside offers are more likely to change firms, so high-seniority individuals will tend to have low productivity and low wages.

Lazear (1986) has proposed a similar "lemon-flavored" model. As in the Harris-Holmstrom model, firms are initially imperfectly informed about worker quality. As a result, some workers are underpaid. Outside firms identify underpaid workers and attempt to hire them. These "raider" firms succeed if they provide a better match for the employee than the incumbent firm. Otherwise, the incumbent firm matches outside offers, and the employee stays. "Raiding" by competitors results in higher wages and higher turnover for high-quality workers, so high-seniority employees tend to be of lower quality and to have lower wages. Low-quality workers remain with the incumbent firm because of the firm's persistent ignorance. In Lazear's model, the negative relationship between seniority and pay depends on employees being valued differently at different firms, since employees only move when they find the raiding firm to be a better "match."

These models justify lower pay for higher seniority because of lower productivity. Although "quality" is a difficult concept to measure, an important product of scholarly activity at research universities is publication. One way to control for productivity is to include variables that measure publication activity in the salary regression models. Table 6 duplicates the analysis found in Tables 2, 3, and 4 but also includes variables to measure the professional publications of the individual.[14] It is clear from Table 6 that a strong negative correlation between salaries and seniority persists, even after controlling for each individual's publication activity. In fact, the estimated penalty for another year of seniority is even greater for the 1973 data when publication variables are included in the regression. For the 1969 analysis, the penalty is greater for seniority intervals up to 10 years. Professors with high seniority do not differ much from individuals with low seniority as far as publication records are concerned. This result suggests that selection by quality is not an adequate explanation for the negative association between salary and seniority.

[14] Publications are measured by a set of interval dummies for (i) number of books, monographs, or manuals published and (ii) number of articles published in academic or professional journals.

IV. A Model of Monopsonistic Discrimination

I propose an alternative "heterogeneity" model to explain the negative slope of the seniority profile. In this model, moving costs drive a wedge between the "market" wage and the wage the employer must pay. Employees with high moving costs will tend to have high seniority and low pay if the employer is able to discriminate.

Consider a simple model of a university that is able to offer a different salary to each current faculty member. Individual faculty members must incur costs to move to another job.[15] The university must meet the market wage to replace a faculty member who rejects the offered salary and chooses to leave. With some uncertainty about an employee's willingness to move, the expected cost to the employer of a specific wage offer would be

$$(1)\quad EC(w_o) = p(w_o, m, w_a)w_o + [1 - p(w_o, m, w_a)]w_a$$

where EC is the expected employment cost (to fill a teaching position), w_o is the wage offered to the individual, w_a is a competitively determined wage rate (what the individual could get if he or she moved), and m is the cost to an individual of moving to a different institution. The probability that an individual will accept the offered wage is $p(\cdot)$, a function of the wage offer, the alternative wage, and the individual's cost of moving. This model assumes that both employer and employee know the employee's alternative salary. The model accommodates no offers or counteroffers.[16]

The only source of uncertainty is that the employer cannot know whether an employee will move at a given offered wage, although the probability of a move is known. A plausible structure for p is

$$(2)\quad p(w_o, m, w_a) = p(w_o + \delta m - w_a)$$

where δ is a discount factor. In this specification, individuals weigh the increased annual income, $w_a - w_o$, against the annualized cost of moving, δm. Those with higher moving cost will be more likely to accept any given wage offer. I assume that the function p is continuous, nonnegative, differentiable, and monotonically increasing.

The wage offer that minimizes the expected employment cost, w_o^* is implicitly defined by the first-order condition:

$$(3)\quad (w_o^* - w_a)p' + p = 0$$

where p' is the derivative of the function p. Since both p and p' are positive, w_o^* must be less than w_a. The sufficient condition for cost minimization is satisfied if

$$(4)\quad 2p' + (w_o^* - w_a)p'' > 0.$$

Taking the total differential of (3), using (2), yields the comparative-statics result:

$$(5)\quad \frac{dw_o^*}{dm} = -\delta\left[\frac{p' + (w_o^* - w_a)p''}{2p' + (w_o^* - w_a)p''}\right] < 0.$$

Thus universities offer lower wages to those who find it most costly to move. Since the expression in brackets in equation (5) must be between 0 and 1, for each dollar increase in the discounted cost of moving the offered wage is reduced by less than one dollar. Therefore, individuals with higher moving costs are more likely to accept the offered salary. Explicitly, the total derivative of equation (2) with respect to m is

$$(6)\quad \frac{dp}{dm} = \delta(p + p')\times\left[1 - \frac{p' + (w_o^* - w_a)p''}{2p' + (w_o^* - w_a)p''}\right] > 0.$$

The expected seniority of an individual is closely related to the value of p. Each year

[15] Most universities are isolated from one another geographically, so it is fair to assume that high moving costs characterize the market for university professors.

[16] In other words, this wage offer is the *last* wage offer that the firm will make. The employee either takes it or leaves.

an individual is offered a salary. If the salary is acceptable, the individual stays; otherwise, he or she leaves. We observe someone with T years of seniority at an institution only if he or she has accepted the wage offer for T consecutive years. Consider all professors who have the same values for w_a and m, (and therefore w_o^* and p). The fraction who move in any year will be $1-p$, so the proportion of this group with no seniority will be $1-p$. Likewise, the proportion with 1 year of seniority will be $(1-p)p$; the proportion with 2 years will be $(1-p)p^2$; and so on. Thus the proportion with T years of seniority will be $p^T(1-p)$. Ignoring the fact that professors have finite careers, expected seniority is

$$(7)\quad E(S) = (1-p)\sum_{\tau=0}^{\infty} \tau p^{\tau} = \frac{p}{1-p}.$$

Expected seniority increases with p, and p increases with m, so professors with high moving costs will have higher average seniority.

Because high values of m not only lead to high values of seniority, but also to low salaries, seniority and wages will be inversely related. In other words, comparing individuals who would obtain the same salary in the market, those with higher current seniority will have lower salaries, *on average*. However, this model does not predict that salaries will fall with seniority for individuals; the negative seniority effect is simply an artifact of heterogeneity. The estimated regression coefficient for seniority is biased downward because the unobserved moving-cost variable, m, is not included. Thus, this model has much in common with the matching model of Jovanovic (1979), although the implications for the relationship between pay and seniority are exactly opposite. Similarly, the recent the papers by Abraham and Farber (1987), Altonji and Shakotko (1987), and Topel (1991) are primarily concerned with sources of heterogeneity that bias the effect of seniority upward. Apparently, the monopsony heterogeneity dominates other sources of heterogeneity for the academic labor market.

Recent papers, such as Topel (1991), have carefully controlled for heterogeneity using panel data. Appropriate data for this type of analysis of the academic labor market are not available. However, since this model identifies moving costs as the source of the sample heterogeneity, if it were possible to include a true measure of moving costs in the salary regression model, then the estimated coefficient for seniority would not be biased downward. Unfortunately, moving costs are difficult to quantify and measure. Not only do they include the costs of physically relocating the household, but they also include psychic costs, such as the pain of separating from friends and family. Furthermore, the available data are limited: it is not even possible to identify city of residence from the national data sets analyzed in this paper.

Marital status, spouse work status, and number of children certainly influence the mobility of families and have often been identified as such in studies of geographical mobility.[17] If the presence of a spouse increases mobility costs, then one would expect, on the basis of the monopsony model presented here, that married men would have higher seniority and lower salaries than single men. However, a simple regression analysis of the 1973 survey[18] shows that although married men do have more seniority than single men, controlling for experience, education, publications, and other variables, a regression analysis of salaries shows that married men have significantly higher salaries than single men. This marriage premium is a well-known phenomenon that is usually explained in terms of heterogeneity. In other words, although married men probably have higher moving costs, they also differ from single men in

[17] For example, Jacob Mincer (1978) explores in some detail the relationship between labor-force participation and geographical mobility for married couples.

[18] The 1973 survey provides the most convenient data for analyzing seniority levels. The other large surveys report seniority in interval form only.

unobservable ways that influence salaries. Thus, marital status fails to provide a test of the monopsony model.

Other types of questions arise from variables of this type, too. For example, we might expect that the moving costs of a married couple would be higher if both spouses have jobs. Indeed, men who have working spouses have higher seniority than men whose spouses do not work, controlling for other factors. Furthermore, married men whose spouses work earn less than married men whose spouses do not work, also as predicted by the monopsony model. On the other hand, family-labor-supply models predict that wives are less likely to work as the income of their husband increases. If the labor participation of the wife is dependent on the income of the husband, then one cannot test the monopsony theory with this variable either. Convincing tests of the monopsony heterogeneity model require better data than are available in the surveys analyzed in this study.

Why do we observe this unexpected relationship between seniority and pay in university labor markets but not in other labor markets? The fact that academic human capital is very general might contribute to this finding. However, even in occupations such as school teaching and accounting, where firm-specific human capital would seem to be unimportant, we see large positive returns to seniority. Furthermore, studies usually find that heterogeneity biases the returns to seniority upward, as most popular theories predict, not downward. There is something peculiar about the institutions of the academic labor market that make the monopsony heterogeneity more important than in other markets. One contributing factor is that distances between universities are relatively large. Except for the largest metropolitan areas, universities are thinly distributed geographically, making moving costs higher for professors than for other occupations. Another aspect of the academic labor market that might increase the importance of monopsony is the individualized nature of pay. For example, Debra A. Barbezat (1989) found positive returns to seniority at colleges represented under collective-bargaining agreements, while returns were negative for employees of other institutions.

V. Summary

Nationwide data from large research-oriented universities show a negative relationship between seniority and salary of professors, even after controlling for individuals' publication records, education, and experience. Salary data collected at three large universities also exhibit a negative return of seniority on salaries. The national data suggest that salaries fall by as much as 0.5 percent per year of seniority. Employment data from individual universities suggest that the decline may be about 1 percent per year of seniority. A professor of average seniority could increase his or her salary by 5–10 percent by moving to a different institution. This is dramatically different than the 25-percent decline that Topel (1991) estimates for the typical American worker.

These findings contradict widely accepted theories of wage determination but are consistent with a heterogeneity model derived from monopsonistic salary discrimination by universities. Individuals with high moving costs receive lower salary offers and have higher seniority than individuals with low moving costs. Differences in publication activity between high-seniority and low-seniority professors are unable to explain the negative returns to seniority, as some other heterogeneity models suggest.

REFERENCES

Abraham, Katherine G. and Farber, Henry S., "Job Duration, Seniority, and Earnings," *American Economic Review*, June 1987, *77*, 278–97.

Altonji, Joseph G. and Shakotko, Robert A., "Do Wages Rise with Job Seniority?" *Review of Economic Studies*, July 1987, *54*, 437–59.

Barbezat, Debra A., "The Effect of Collective Bargaining on Salaries in Higher Education," *Industrial and Labor Relations Review*, April 1989, *42*, 443–55.

Bayer, Alan E., "Teaching Faculty in Aca-

deme: 1972–73, *ACE Research Reports*, August 1973, *8*, 1–68.

Becker, Gary S. and Stigler, George, "Law Enforcement, Malfeasance, and Compensation of Enforcers," *Journal of Legal Studies*, January 1974, *3*, 1–18.

Blum, Debra E., "Job Market Begins Predicted Turnaround, Survey Indicates; Many Colleges Finding It Difficult to Fill Faculty Vacancies," *Chronicle of Higher Education*, 25 June 1990, *36*, A1.

Harris, Milton and Holmstrom, Bengt, "Theory of Wage Dynamics," *Review of Economic Studies*, July 1982, *72*, 716–24.

Hoffman, Emily, "Faculty Salaries: Is There Discrimination by Sex, Race, and Discipline. Additional Evidence," *American Economic Review*, March 1976, *72*, 196–8.

Jovanovic, Boyan, "Job Matching and the Theory of Turnover," *Journal of Political Economy*, October 1979, *87*, 972–90.

Lazear, Edward P., "Agency, Earnings Profiles, Productivity and Hours Restrictions," *American Economic Review*, September 1981, *41*, 606–20.

_____, "Raids and Offer Matching," *Research in Labor Economics*, 1986, *8* (Part A), 141–65.

Megdal, Sharon B. and Ransom, Michael R., "Longitudinal Changes in Salary at a Large Public University: What Response to Equal Pay Legislation?" *American Economic Review*, May 1985 (*Papers and Proceedings*), *75*, 271–4.

Mincer, Jacob, *Schooling, Experience, and Earnings*, New York: Columbia University Press, 1974.

_____, "Family Migration Decisions," *Journal of Political Economy*, October 1978, *86*, 749–73.

Oi, Walter, "Labor as a Quasi-fixed Factor," *Journal of Political Economy*, December 1962, *70*, 538–55.

Salop, Joanne and Salop, Steven, "Self Selection and Turnover in the Labor Market," *Quarterly Journal of Economics*, November 1976, *90*, 619–27.

Stiglitz, Joseph, "The Causes and Consequences of the Dependence of Quality on Price," *Journal of Economic Literature*, March 1987, *25*, 1–48.

Topel, Robert, "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority," *Journal of Political Economy*, February 1991, *99*, 145–76.

Bureau of National Affairs, *Government Employee Relations Report*, March 1991, *29*, 259.

Carnegie Commission, *Carnegie Commission National Survey of Higher Education: Faculty Study, 1969*, ICPSR Edition, Ann Arbor, MI: Interuniversity Consortium for Political and Social Research, 1978.

U.S. Department of Commerce, *Current Population Survey, May 1988: Survey of Employee Benefits Technical Documentation*, Washington, DC: Bureau of the Census, 1989.

The Decline in Black Teenage Labor-Force Participation in the South, 1900–1970: The Role of Schooling

By Robert A. Margo and T. Aldrich Finegan*

In 1950, approximately 61 percent of Southern black males aged 16–19 were participating in the labor force. By 1970 their participation rate had declined to 34 percent, or by 27 percentage points in two decades. Perhaps the most widely accepted explanation of the post-1950 decline in black teen participation emphasizes shifts in labor demand (John Cogan, 1982; Gavin Wright, 1986; see also Richard Day, 1967). Between 1950 and 1960 the introduction of the mechanical harvester in cotton agriculture substantially reduced the demand for the labor of black male teenagers, who were concentrated in cotton production. Black teens displaced from Southern cotton farms did not find ready work in the South's nonfarm economy. Their failure to find nonfarm jobs, according to Cogan, was caused by increases in the coverage of the federal minimum wage in the 1960's. The combination of cotton mechanization and the minimum wage drastically reduced employment opportunities for black male teens in the South, resulting in the sharp decline in their labor-force participation between 1950 and 1970.

We argue that the post-1950 effects of cotton mechanization and the minimum wage have been overrated. A portion of the post-1950 decline in black teen participation was the continuation of a downward trend whose origins can be dated at least to the early 1900's. The secular downward trend in participation was the mirror image of a long-term upward trend in black school enrollment. Because the factors producing the long-run increase in schooling had not yet run their course by mid-century, a decline in black teen participation in the South between 1950 and 1970 would have happened anyway, even if cotton mechanization and changes in the level and coverage of the minimum wage had been delayed until the 1970's.

I. Estimates of Labor-Force Participation Rates

Census data provide the basis for estimating pre-1950 labor-force participation rates of black male teens (ages 16–19) in the South. Our estimates are shown in Figure 1, along with post-1950 participation rates.[1] In 1900, 86 percent were participating in the labor force. Their participation rate fell by 10 percentage points between 1900 and 1920, and by an additional 15 percentage points between 1920 and 1950—or an average rate of decline of almost 5 percentage points per decade between 1900 and 1950. Fully 48 percent of the decline in participation between 1900 and 1970 took place before 1950.

Cogan (1982) attributed the post-1950 decline in black teen participation to the negative effects of mechanization on agricultural labor demand, primarily in cotton agriculture, and to the effects of the minimum

*Department of Economics, Vanderbilt University, Nashville, TN, 37235. We are grateful to Lee Alston, Jeremy Atack, Stanley Engerman, Keith Inlanfeldt, Thomas Kane, Frank Sloan, Gavin Wright, and seminar participants at Georgia State University, the University of Illinois, the National Bureau of Economic Research, Northwestern University, the University of Wisconsin, and the ASSA-Cliometrics Society meetings for helpful comments.

[1]We do not estimate an overall participation rate for 1910 because the census is thought to have overcounted participation among farm children in that year (John Durand, 1948).

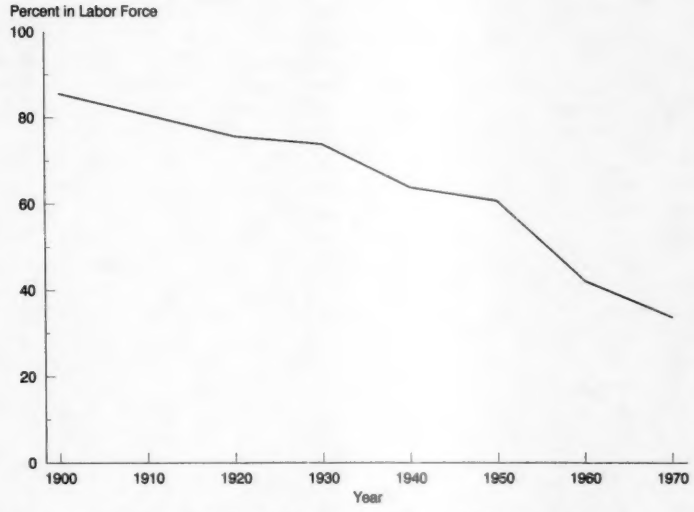


FIGURE 1. LABOR-FORCE PARTICIPATION RATES OF SOUTHERN BLACK MALES, AGED 16–19

Note: The points in the graph are: 1900, 85.6 percent; 1920, 75.7 percent; 1930, 73.9 percent; 1940, 63.9 percent; 1950, 60.8 percent; 1960, 42.2 percent; 1970, 33.8 percent.
Source: See Data Appendix.

wage on labor demand off the farm.[2] Between 1950 and 1970 the number of man-hours required to produce a bale of cotton declined sharply after the widespread diffusion of the mechanical cotton picker. Because cotton production remained constant over the period, the decline in labor requirements translated into a one-for-one reduction in labor demand. Cogan argued that virtually the entire decline in employment of black teens in the South between 1950 and 1970 could be explained by the decline in their employment in agriculture. Although some black teens found jobs in manufacturing, growth in their employment in retail trade slowed considerably in the 1960's, precisely when coverage of the minimum wage was extended to the industry.

Whatever the merits of Cogan's argument, Figure 1 demonstrates that participation rates of black male teens were already falling between 1900 and 1950. Had post-1950's participation rates continued to decline at the average decadal rate experienced between 1900 and 1950 (4.96 percentage points), the predicted participation rate in 1970 would have been 51 percent. The difference between this predicted 1970 rate and the actual 1950 rate accounts for 37 percent (= 9.9/27) of the decline in participation between 1950 and 1970. If the predicted 1970 rate is calculated using the average (decadal) rate of decline between 1930 and 1950, the percentage explained is 49 percent. The decline in black teen participation accelerated after 1950, but a post-1950 reduction in participation could have

[2]Cogan analyzed the decline in the employment-to-population ratio, not the participation rate. This ratio fell by 27 percentage points between 1950 and 1970, the same as the decline in the participation rate (Cogan, 1982 p. 623). We focus on participation rates because employment ratios cannot be constructed for census years prior to 1930.

been predicted on the basis of the pre-1950 trend. Indeed, such a prediction was made in 1948.[3]

Although the problems of designing a reliable mechanical cotton picker had largely been solved by the late 1920's, marketing of the machine was delayed until the early 1940's, and the picker did not diffuse until after 1950 (Cogan, 1982 p. 627; Warren Whatley, 1987 pp. 66–7). Cotton mechanization, therefore, cannot explain the pre-1950 trend. Might some other type of agricultural mechanization or the minimum wage account for the pre-1950 decline in participation?

Consider the decline in participation before 1930. The minimum wage cannot possibly explain the pre-1930 decline in participation because minimum-wage legislation was first enacted in 1938.[4] The principal type of mechanization occurring in American agriculture prior to World War II was the diffusion of tractors (Moses Musoke, 1981). The diffusion of tractors was far slower in the South than elsewhere in the country. In 1930, only 4 percent of Southern farms used tractors (U.S. Bureau of the Census, 1952a p. 226). Yet the participation rate of black teens fell by 12 percentage points between 1900 and 1920, before virtually any tractors had been introduced into Southern agriculture (U.S. Bureau of the Census, 1952a p. 226).

Tractorization or the minimum wage, however, might account for some of the decline in participation between 1930 and 1950. The proportion of Southern farms using tractors rose to 26 percent by mid-century (U.S. Bureau of the Census, 1952a p. 226). Although tractors did not substantially reduce labor requirements in cotton compared with the mechanical cotton picker, tractorization was associated with a decrease in sharecropping and other forms of tenancy and with a redistribution of cotton production to the Southwest and California (Musoke, 1981; Whatley, 1983; Lee Alston, 1985). Because tenant labor was frequently supplied in family groups, the decline in tenancy may have reduced the agricultural demand for black teen labor, as might the spatial reallocation of cotton production. Estimates of industry coverage rates of the minimum wage indicate that about 20 percent of employed Southern black teens were in covered employment in 1950.[5]

To evaluate the pre-1950 impact of tractorization and the minimum wage, we estimated a regression using data for 16 Southern states from 1930 and 1950. The dependent variable is the change in the black teen participation rate (dLFPR) between 1930 and 1950. The independent variables are the change in proportion of farms using tractors (dTRACTORS; from U.S. Bureau of the Census, 1952a p. 226) and the coverage rate of the minimum wage in 1950 (dCOVERAGE). Observations were weighted by the average of population counts in 1930 and 1950. The results are:

$$d\text{LFPR} = \underset{(1.356)}{-0.105} - \underset{(0.265)}{0.042}\, d\text{TRACTORS} - \underset{(0.043)}{0.014}\, d\text{COVERAGE}$$

where absolute values of t statistics are shown in parentheses. The regression completely fails to explain the variance in the dependent variable ($\bar{R}^2 = -0.16$), although the coefficients of the tractorization and

[3]Durand (1948 pp. 66, 254) predicted a decline of 7.6-percentage points in the participation rate of nonwhite males aged 14–19 between 1950 and 1960.

[4]Wright (1986) argued that the National Industrial Recovery Act (NIRA) was the first federal minimum-wage legislation. Our argument about the pre-1930 decline in participation is unaffected, however, because the NIRA was passed in 1933.

[5]See the Data Appendix on the construction of the coverage rates. The average coverage rate was low because most jobs held by black teens were in agriculture or the service sector, and most jobs in these industries were exempt from the minimum wage in 1950.

coverage variables are both negative.[6] The decline in participation between 1930 and 1950 was 13.1 percentage points (Fig. 1). Treating the coefficients as the best point estimates of their true effects, tractorization and the minimum wage can explain about 20 percent $[=1-(0.105/0.131)]$ of the 1930–1950 decline in participation.[7] Variations across states in the rate of decline were essentially unrelated to the diffusion of tractors or the extent of coverage of the minimum wage in 1950.

The regression and Figure 1 might be criticized, however, because of alterations in labor-force statistics over the period. In 1940 the definition of the labor force changed to the labor-force-week concept. As Durand (1948) pointed out, the pre-1940 participation rates are not, strictly speaking, comparable with the post-1940 estimates (see also Stanley Lebergott, 1964). Two responses to this criticism can be made. First, changes in the definition of the labor force have little bearing on the finding of a pre-1940 decline in participation, because the gainful-worker definition was used in 1900 and 1930.[8] Second, Durand (1948) calculated adjustment ratios in an attempt to make the pre- and post-1940 participation rates comparable. The force of Durand's adjustments was to reduce the size of the labor force prior to 1940. Had the 1940 census question been asked in 1900 and 1930, participation rates in both years would have been lower.

Data limitations prevented Durand (and us) from estimating decade-specific adjustment ratios for Southern black male teens, but Durand did estimate an adjustment ratio for all males aged 14–19 in 1930: 0.9756 (Durand, 1948 p. 199). Multiplying our estimate of the 1930 participation rate for Southern black male teens by Durand's adjustment ratio produces an adjusted participation rate of 72.1 percent, compared with the measured participation rate of 73.9 percent. Even if we assume an adjustment ratio equal to the lowest ratio estimated by Durand for any group of males (0.8953 for males aged 75 and over), there still was a decline in black teen participation between 1930 and 1950 (Durand, 1948 p. 199).[9] It is unlikely that adjusting for definitional changes would alter the finding of a decline in black teen participation in the South before 1950.

[6]We reestimated the regression with the coverage rate defined to be the percentage in covered employment multiplied by the ratio of the minimum wage in 1950 to the average hourly wage in manufacturing in the state. The substantive results were unchanged.

[7]The impact of the minimum wage may be understated because we are examining participation and not employment. The minimum wage can increase participation through the added-worker effect, while simultaneously reducing employment. Although it is not possible to measure the change in the employment-to-population ratio for Southern black males aged 16–19 between 1930 and 1950, it is possible to calculate (from U.S. Bureau of the Census, 1931) a proxy for the unemployment rate of 15–19-year-olds in 1930 (the number of unemployed in the census week divided by the count of gainful workers). Since the unemployment rate in 1950 of 15–19-year-olds is known (U.S. Bureau of the Census, 1953), we can estimate the decline in the employment-to-population ratio for 15–19-year-olds and compare it with the decline in participation. The employment-to-population ratio fell 14.9 percentage points between 1930 and 1950 for 15–19-year-olds, compared with a decline in participation of 13.1 percentage points. The decline in employment was only slightly larger than the decline in participation, consistent with our finding that the effect of the minimum wage prior to 1950 was small.

[8]It might be argued that the measured decline in participation between 1900 and 1920 was overstated because the 1920 census was taken in January, a seasonal bottom in teen employment, while the 1900 census was taken in June, a seasonal peak. Durand (1948 p. 194), however, argued that "the proportion of youthful seasonal workers reported as having gainful occupations was not very much affected by their employment status at the time the census was taken." In addition, Paul Osterman (1980 p. 174) claimed that instructions to census enumerators caused some youth employment to be missed in 1900, which would understate the measured decline in black teen participation between 1900 and 1920.

[9]Applying the adjustment ratio for males aged 75 and over, the estimated 1930 participation rate is 66.1 percent, 5 percentage points higher than the 1950 rate.

TABLE 1—CIVILIAN LABOR-FORCE PARTICIPATION OF SOUTHERN BLACK MALE TEENS, AGED 16–19: BY SCHOOL ENROLLMENT STATUS, 1900–1970

Year	Percentage in farm sector	Percentage enrolled in school	Percentage in labor force		
			Dropouts	In school	Total
Farm:					
1900	59.5	17.7	90.2	66.7	86.0
1930[a]		29.2	88.5	55.1	78.7
1950	40.1	38.4	89.6	46.3	72.8
1960	18.2	57.3	73.6	26.1	46.4
1970	4.7	70.9	66.4	12.2	28.0
Nonfarm:					
1900[b]		11.9	92.8	44.0	86.9
1950		45.5	80.1	22.0	52.7
1960		60.3	68.9	23.3	41.3
1970		66.0	61.8	20.9	34.1

Notes: Civilian labor-force participation rates for all black male teens shown in the "total" column were calculated from regional data; rates for enrollment subsets were calculated from state data with some missing observations (see the Data Appendix). Hence, the total rates differ slightly from weighted averages of the within-enrollment-status rates.

Sources: 1900, authors' calculations from 1900 and 1910 census public-use samples, see note b below; 1930, U.S. Bureau of the Census (1935 p. 226); 1950, 1960, and 1970, see the Data Appendix.

[a]Participation rates for the farm population by enrollment status are available for the age group 16–19 for the entire United States, but not the South. However, 98 percent of the rural black farm population resided in the South in 1930 (U.S. Bureau of the Census, 1935 p. 51). The fraction of Southern black male teens in 1930 in the rural farm sector cannot be calculated from published census data.

[b]Nonfarm 1900 combines 1900 and 1910 public-use samples (see Population Studies Center, 1989) because sample size in 1900 is too small to permit reliable disaggregation by enrollment status at the regional level.

II. Labor-Force Participation and Schooling: The Impact of Age-in-Grade Retardation

If the minimum wage, tractorization, or biases in census data cannot explain the majority of the pre-1950 decline in black teen participation, which factors can? Some clues are suggested by Table 1, which gives participation rates by farm–nonfarm and school enrollment status, along with the percentage on farms and the percentage enrolled in school.

At the turn of the century over 90 percent of Southern black male teen "dropouts" (i.e., those not enrolled in school) were participating in the labor force. Participation rates of in-school teens were lower than participation rates of dropouts but were still relatively high by comparison with the rates experienced after World War II. Despite a lower participation rate among in-school nonfarm teens, the overall participation rate did not differ between the farm and nonfarm population, because the school enrollment rate was actually higher among farm teens.

In 1950, the participation rate of farm teen dropouts was approximately the same as in 1900. The participation rate of in-school farm teens, however, declined by 20 percentage points, and the proportion in school increased by 21 percentage points between 1900 and 1950. Among nonfarm teens, the percentage enrolled in school increased by 34 percentage points, and the participation rate of in-school black teens declined by 22 percentage points between 1900 and 1950. Unlike the farm population, however, the participation rate of nonfarm

TABLE 2—HYPOTHETICAL PARTICIPATION RATES: SOUTHERN BLACK MALE TEENS

Rate	Participation rate (percentage)		
	Farm	Nonfarm	Total
A. Hypothetical rates for 1900 given:			
1950 enrollment rates	81.2	70.6	76.9
(Percentage explained)			(38.3)
1950 enrollment rates and 1950 in-school labor-force participation rates	73.3	60.6	68.2
(Percentage explained)			(73.4)
1950 not-in-school labor-force participation rates	85.5	76.4	81.8
(Percentage explained)			(18.5)
B. Hypothetical rates for 1950 given:			
1970 enrollment rates	58.9	42.1	48.8
(Percentage explained)			(48.9)
1970 enrollment rates and 1970 in-school labor-force participation rates	34.7	41.3	38.7
(Percentage explained)			(87.4)
1970 not-in-school labor-force participation rates	58.8	45.2	50.6
(Percentage explained)			(42.0)

Notes: Figures give counterfactual labor-force participation rates under various assumptions. For example, row 2 of panel A calculates participation rates in 1900 assuming 1950 school enrollment rates, 1950 in-school labor-force participation rates, and no other changes between 1900 and 1950. The numbers in parentheses are the percentages of decline in actual labor-force participation rates between 1900 and 1950 or between 1950 and 1970 explained by the difference between the actual and counterfactual labor-force participation rates.
Source: Calculated from Table 1.

teen dropouts decreased by almost 13 percentage points. Nonfarm teens experienced a decline in participation over 2.5 times as large as the decline in participation among farm teens because of the larger rise in nonfarm school enrollment and the fall in participation among nonfarm teen dropouts.

School enrollment rates continued to climb between 1950 and 1970. Among farm teens in school, participation rates fell by 34 percentage points between 1950 and 1970, but the participation rate of nonfarm teens enrolled in school remained stable at slightly over 20 percent. In contrast to the pre-1950 period, participation rates of farm teen dropouts fell by 23 percentage points between 1950 and 1970. Participation rates of nonfarm teen dropouts continued to fall as they had before 1950, but at a faster rate, with most of the decline concentrated in the 1950's. Thus, reversing the patterns for the pre-1950 era, from 1950 to 1970 the overall participation rate for farm teens fell more than twice as much as the rate for nonfarm teens.[10]

[10]See Robert Mare and Christopher Winship (1984) for a detailed analysis of post-1960 interactions between black school attendance and labor-force participation, based on CPS data.

Table 1 suggests that increasing rates of school enrollment and reduced labor-force activity among black teens in school were proximate causes of the pre-1950 decline in participation. The hypothesis is confirmed in Table 2, which computes participation rates for 1900 and 1950 under various assumptions about school enrollment rates and participation conditional on enrollment status. Had enrollment rates (farm and nonfarm) in 1900 been equal to their 1950 values, the overall participation rate in 1900 would have been 77 percent. Had enrollment rates and participation rates of in-school black teens in 1900 equaled their 1950 values, the overall participation rate would have been 68.2 percent. The difference between the actual 1900 participation rate and the latter counterfactual rate accounts for approximately three-quarters of the decline in participation between 1900 and 1950. Changes in participation among dropouts were less important, explaining about 18 percent of the decrease in participation before 1950.

Increasing school enrollments and decreasing participation rates of in-school teens continued to be important after 1950, but the significance of changes in participation among dropouts was greater after 1950 than before. Because nonfarm participation rates were lower than farm participation rates, the shift of the population out of agriculture between 1950 and 1970 (35 percentage points) also helped to lower the overall participation rate.[11]

The sharp decline after 1950 in participation among farm teen dropouts is consistent with Cogan's (1982) argument that cotton mechanization reduced employment opportunities for black teens in Southern agriculture. The post-1950 shift of black labor out of agriculture, the increase in school enrollment, and the decrease in participation among in-school teens might also be attributed to cotton mechanization, except that downward trends in these variables began long before 1950.[12] What needs to be explained are changes in school enrollments and the high participation rate of black teens in school prior to 1950.

We hypothesize that changes before 1950 in school enrollment rates and in participation rates of black teens in school were generated, in part, by rising expectations of black parents that their children complete more years of schooling before entering the labor force permanently. At the turn of the century black parents in the South were poor, poorly educated, and heavily rural. Poverty, adult illiteracy, the perception that the returns to schooling were low in agriculture, the value of teen labor on family farms, and poor-quality schools conspired to keep the typical black child out of school at later ages. In 1910, for example, the peak age of black school enrollment in the South was age 11 (71 percent), and enrollment rates fell off sharply at age 14 (U.S. Bureau of the Census, 1918 p. 384). The upshot was that Southern blacks educated in the late 19th and early 20th centuries finished very few years of schooling before entering the labor force: an average of about five years for the birth cohorts from 1886 to 1905 (James Smith, 1986 p. 1227).[13]

That the number of years of schooling desired was so low can account for the low enrollment rate of black teens in 1900, but it cannot explain why the labor-force participation of in-school black teens was so high in the early 20th century. A child whose

[11] Interaction effects were also important after 1950. The fraction of the post-1950 decline in the overall participation rate attributed to increased enrollments and lower in-school participation rates (87 percent), when added to the fraction attributed to lower participation of dropouts (42 percent), exceeds 100 percent.

[12] In addition, if technical change in Southern agriculture (other than tractorization) were solely responsible for the pre-1950 trends in enrollments, percentage on farms, and participation rates of in-school farm teens, it is difficult to explain why the participation rate of farm teen dropouts did not decline until after 1950.

[13] Smith's figure pertains to the Southern black population. For black males it is biased upward because black female children were more likely to attend school than black male children around the turn of the century; it is biased upward more generally because the census overstated black educational attainment at the turn of the century (Margo, 1990 Ch. 2).

education stopped at the sixth grade could, in theory, finish his schooling by age 12 or 13 and then enter the labor force. In fact, a significant fraction of Southern black children suffered from high rates of age-in-grade retardation and simply did not complete a given number of years of schooling in the customary amount of time (Finis Welch, 1973). According to the 1940 census, for example, 47 percent of Southern black males enrolled in school at age 12 were attending the fourth grade or a lower grade (U.S. Bureau of the Census, 1943a p. 126). Although age-in-grade retardation in the preteen years had many causes, there is little doubt that a major factor was the poor quality of elementary schools attended by Southern black children (Horace Mann Bond, 1934 pp. 344–56; Finis Welch, 1973).

As long as the parental demand for child's schooling was low, age-in-grade retardation at an early age was not a binding constraint. Completion of a given number of years of schooling could be spread over several years and be readily combined with labor-force participation at later ages. Once desired years of schooling reached modern norms, however, enrollment rates at later ages increased, and participation of in-school teens fell. A black teen struggling to finish his schooling with the sixth grade could work and still expect to achieve his goal before leaving home. However, a black teen whose parents wished him to complete high school would, under normal circumstances, graduate at age 18. He could ill afford to fall several years behind, at an early or a later age, and his parents could best accommodate their wishes by keeping him out of the labor force as much as possible during his teenage years.

Over time, more black parents sought to have their children complete higher levels of schooling, while improvements in the quality of black elementary schools, particularly the length of the school year, lessened the extent of early age-in-grade retardation.[14] The reasons for the increased parental demand are not fully understood, but reductions in adult illiteracy and poverty clearly played important roles.[15] Cross-sectional analysis of data from the early-20th-century South reveals significant positive effects of adult literacy and the

[14] Using state-level data for Southern states in 1940 and 1950, we estimated a regression of the change in the retardation rate at age 12 (the percentage of enrolled black 12-year olds attending the fourth grade or some lower grade) between 1940 and 1950. The independent variables were the change in the length of the school year in days, the change in the pupil-to-teacher ratio, and the change in the average daily attendance rate. Delaware was omitted because the retardation rate was not available for 1950. Data were weighted by the population counts (enrolled 12-year-olds) in each state. The retardation rate was calculated from U.S. Bureau of the Census (1943b) and U.S. Bureau of the Census (1952b). The sources for the length of the school year and the pupil-to-teacher ratio were Federal Security Agency (1947) and U.S. Department of Health, Education, and Welfare (1954). The coefficient for the length of the school year was -0.0029 (t statistic $= -1.85$). The coefficients of the pupil-to-teacher ratio and the daily attendance rate had the wrong signs (negative and positive, respectively), but neither was statistically significant. The retardation rate at age 12 fell by 6.9 percentage points between 1940 and 1950, while the average length of the school year rose by 17.4 days. Based on the regression, we estimate that 73 percent ($= -0.0029 \times 17.4/0.069$) of the decrease in the retardation rate between 1940 and 1950 is explained by the increase in the length of the school year.

[15] Three other factors are worth noting. A significant part of the returns to schooling for Southern blacks was migration to the North, where real wages were higher. Better-educated blacks were more likely to leave the South than less-educated blacks (Margo, 1990 Ch. 7). Migration to the North became more feasible with the establishment of black communities in Northern cities (Wright, 1986), and awareness of the economic benefits of migration may have prompted some black parents to keep their children in school at later ages. Second, violations of the separate-but-equal doctrine were much worse at the high-school level than at the elementary level in the South in the early 20th century. Few black teenagers in the rural South in 1910 could have attended a local public high school even if they wanted to, because the number of such schools was extremely small (Margo, 1990 p. 20). Provision of black high schools became more common, however, after World War I. Third, there is evidence of "catch-up" behavior among Southern black parents; that is, holding constant family background, school quality, and local labor-market characteristics, black parents were more likely to send their children to school than were white parents (Margo, 1990 pp. 79–80).

TABLE 3—ENROLLMENT AND LABOR-FORCE PARTICIPATION REGRESSIONS: SOUTHERN BLACK MALE TEENS, 1950

Independent variable	Enrollment rate			Labor-force participation of in-school teens		
	WLS	WLS	2SLS	WLS	WLS	2SLS
Constant	0.508	0.574	0.580	0.239	0.078	0.066
	(23.989)	(13.156)	(10.558)	(8.372)	(1.592)	(1.538)
Elementary enrollment ratio		−0.277	−0.298		0.662	0.713
		(−1.739)	(−1.421)		(3.692)	(2.827)
Rural nonfarm	−0.094	−0.050	−0.047	0.072	−0.103	−0.111
	(−5.337)	(−1.667)	(−1.268)	(0.030)	(−3.043)	(−2.489)
Farm	−0.103	−0.024	−0.018	0.227	0.038	0.023
	(−6.728)	(−0.495)	(−0.289)	(10.989)	(0.703)	(0.316)
Mean value of dependent variable:	0.43			0.33		
$\bar{R}^2$:	0.77	0.78	0.78	0.91	0.91	0.90
N:	40			40		

Notes: WLS denotes weighted least squares; 2SLS denotes weighted two-stage least squares. Numbers in parentheses are t statistics.
Source: See Data Appendix and text.

household head's occupational and home-ownership status (proxies for income and wealth) on the school attendance of black children (Margo, 1990 pp. 13, 77).[16]

The regressions in Table 3 provide a test of our hypothesis. The data are within-state aggregates for farm, rural nonfarm, and urban black teens in the South in 1950. The dependent variables are the school enrollment rate and the labor-force participation rate of in-school black teens. The key independent variable is the elementary enrollment ratio: the percentage of total school enrollment among black male teens accounted for by the elementary grades. Dummy variables are included for states and for rural nonfarm and urban status to control for a variety of factors that may have influenced participation and school enrollment but which cannot be measured directly.[17] The coefficients in the columns labeled WLS are estimated using weighted least squares; the weights are population counts. Data are drawn from the state volumes of the 1950 census (see the Data Appendix).

As our hypothesis predicts, the coefficient of the elementary enrollment ratio is negative in the enrollment regression and positive in the labor-force participation regression. Every 10-percentage-point decrease in the elementary enrollment ratio increases the enrollment rate by 2.8 percentage points and decreases the labor-force participation rate of in-school teens by 6.6 percentage points. Controlling for the elementary ratio substantially accounts for the large differences in labor-force participation and school enrollment between the farm, rural nonfarm, and urban populations evident in the regressions excluding this variable.

[16]Compulsory-schooling legislation also raised black school attendance, but the effect was small (see Margo, 1990 p. 13).

[17]The inclusion of state and within-state sectoral dummies is analogous to a fixed-effects regression. In our application the fixed effects are states, and the within-state dummies correspond to time dummies.

An objection to the weighted-least-squares estimates is that the elementary enrollment ratio might be correlated with the error term. The number of teens enrolled in school is the denominator of the elementary enrollment ratio and also the numerator of the enrollment rate and the denominator of the labor-force participation rate. Measurement error in enrollments would cause a spurious negative correlation between the enrollment rate and the elementary enrollment ratio and a spurious positive correlation between the labor-force participation rate and the elementary enrollment ratio. Because the elementary enrollment ratio was constructed from the state volumes of the 1950 census, which drew upon a 20-percent random sample of the population, there is reason to suspect measurement error.

The regressions labeled 2SLS in Table 3 were estimated using weighted two-stage least squares. The instruments for the elementary enrollment ratio were the median years of schooling of males aged 45–54, median household income (education of adult males and household income are proxies for parental demand), the retardation rate at age 10 (the fraction of 10-year-olds enrolled in the third grade or lower, a proxy for school quality), and the state and sectoral dummies. Although the significance level of the elementary enrollment ratio in the enrollment regression decreases, the two-stage coefficients have the same signs and are slightly larger than the weighted-least-squares coefficients.[18] We conclude that our substantive findings are unchanged, allowing for the possible endogeneity of the elementary enrollment ratio.

We use the weighted-least-squares regression coefficients to compute counterfactual enrollment and participation rates based on different assumptions about the value of the elementary enrollment ratio. In 1950, the value of the elementary ratio for Southern black male teens was 0.376 (U.S. Bureau of the Census, 1953 p. 350). Had the ratio in 1950 been equal to its value in 1970 (8.5 percent; calculated from U.S. Bureau of the Census [1973a p. 1116]) the estimated participation rate would have been 46.1 percent.[19] The difference between the actual and counterfactual 1950 participation rates accounts for 54 percent (= 14.7/27) of the decline in black teen participation between 1950 and 1970.

In 1940, the elementary enrollment ratio for Southern black male teens was 51.2 percent (U.S. Bureau of the Census, 1943a p. 127). Comparable figures for the pre-1940 period are not available, but it is possible to estimate the ratio from information on the age structure of enrollments and the number of pupils enrolled in high school. Using these data we estimate that the ratio was 58.9 percent in 1930.[20] Thus the percentage of black teens enrolled in

[18]We also experimented with different combinations of instruments, including the median years of schooling of females aged 45–54 (in addition to, or instead of, median household income) and the retardation rate at age 9 (instead of the retardation rate at age 10). The significance level of the two-stage coefficient of the elementary ratio in the enrollment regression was sensitive to the choice of instruments, but the sign of the coefficient was always negative, and the magnitude of the coefficient was always close to the value reported in Table 3. The significance level, sign, and magnitude of the coefficient of the elementary ratio in the participation regression were robust to the choice of instruments.

[19]Based on the weighted-least-squares coefficients in Table 3, we calculate the average value of the enrollment and labor-force participation rates for farm, rural nonfarm, and urban sectors, setting the elementary enrollment ratio equal to 0.085 (the value for 1970). We assume that the sector-specific labor-force participation rates for black teen dropouts equal their actual 1950 values. Using the 1950 proportions of black teens in each sector as weights, we calculate the overall participation rate.

[20]We calculated the elementary enrollment ratio in 1930 by subtracting the number of Southern black pupils enrolled in high school (from U.S. Department of the Interior, 1932) from the total number of Southern blacks, aged 16–19, attending school in 1930 (from U.S. Bureau of the Census, 1935 p. 213). The estimate is biased downward for males because the data on enrollments by grade are not gender-specific, and black female children were more likely to attend school than black male children. The ratio is also biased downward because we assume that no 15-year-olds were enrolled in high school, when some were. The latter bias is offset, however, by the implicit assumption that no 19-year-olds were enrolled in college.

the high-school grades had been increasing for at least two decades prior to 1950, producing a decline in labor-force participation that could have been expected to continue after 1950.[21]

Racial and regional differences in black teen participation can also be partly attributed to differences in the elementary enrollment ratio. The ratio for Southern white male teens was 11.8 percent in 1950 (U.S. Bureau of the Census, 1953 p. 350). Had there been no racial difference in the ratio in the South in 1950, we estimate that 48 percent of black male teens would have been participating in the labor force, 7 percentage points *less* than the participation rate of Southern white male teens in 1950. The fact that black teens were less likely to participate in the labor market than white teens, holding the elementary enrollment ratio constant, suggests that racial discrimination may have played a role in limiting labor-market opportunities in the South before 1950 (Margo, 1990 Ch. 6).

In 1950 the elementary enrollment ratio for non-Southern black teens was 15.7 percent (calculated from U.S. Bureau of the Census [1953 pp. 345, 348, 354]). Had the ratio for Southern black teens equaled the non-Southern value, their participation rate is predicted to have been 50 percent. Since the participation rate of non-Southern black teens was 40.4 percent in 1950 (calculated from U.S. Bureau of the Census [1953 pp. 366–73]), the regional difference in the elementary enrollment ratio explains 53 percent ($=10.8/20.4$) of the regional difference in participation rates.

III. Conclusion

The decline in labor-force participation among black male teens in the South between 1950 and 1970 has frequently been attributed to post-1950 effects of cotton mechanization and the minimum wage. We have shown, however, that the post-1950 decline in participation was partly the continuation of a secular trend. Parental demand was an important factor behind the secular trend: as black parents desired higher levels of educational attainment for their children than had been the norm early in the century, low rates of teenage school enrollment and high rates of labor-force participation among in-school teens were no longer feasible.

The limitations of our argument should be recognized. Increased parental demand is not responsible for the post-1950 decline in participation among farm and nonfarm dropouts, or the pre-1950 decline in participation among nonfarm dropouts.[22] Parental demand cannot fully account for racial and regional differences in participation in 1950. Productivity growth in the Southern economy before 1950, including agriculture, raised the incomes of black parents and thus contributed indirectly to the secular trend in participation. There is compelling evidence that the minimum wage had severe disemployment effects on black teens in certain Southern industries (e.g., sawmills) after 1950 (Cogan, 1982). The important point, however, is that debate over the causes of the post-1950 decline in black teen participation should no longer ignore the pre-1950 trend and its causes.

Data Appendix

All data are available from Robert A. Margo on a floppy disk.

[21] If the value of the elementary ratio in 1950 had equaled its estimated value in 1930, the predicted participation rate in 1950 is 69.9 percent. The decline in participation between 1930 and 1950 was 13.1 percentage points. Thus the decrease in the elementary enrollment ratio between 1930 and 1950 explains 70 percent ($=9.1/13.1$) of the decline in participation between 1930 and 1950.

[22] The precise timing of the pre-1950 decline in participation among nonfarm dropouts cannot be determined from the published census volumes. For the country as a whole, the participation rate of nonfarm black teen dropouts was 89.6 percent in 1900 (computed from the public-use sample of the 1900 census [Center for Studies in Demography and Ecology, 1980]) and 83.1 percent in 1930 (U.S. Bureau of the Census, 1935 p. 226). This suggests that the minimum wage cannot be fully responsible for the pre-1950 decline in participation among nonfarm dropouts, because some of the decline occurred before the passage of minimum-wage legislation.

Labor-Force Participation Rates.—Prior to 1940 the census identified persons aged 10 and over who reported a gainful occupation. Participation rates for 1900, 1920, and 1930 are the fractions of Southern black males, aged 16–19, reporting a gainful occupation. The participation rate for 1900 (Fig. 1) was calculated from the public-use sample of the 1900 census (Center for Studies in Demography and Ecology, 1980).[23] Participation rates for 1920 and 1930 were calculated from state-level census data on population and occupations by age (U.S. Bureau of the Census, 1922 pp. 189–285; 1923 pp. 446–51, 515–18; 1933 table 9; 1935 pp. 108–9). The participation rates for 1900, 1920, and 1930 pertain to blacks.

Participation rates for 1940–1960 pertain to the labor-force activity of nonwhite male teens in the South during the census week. For 1970 the participation rate pertains to black teens exclusively. Virtually no bias is introduced by the distinction between nonwhites and blacks because blacks accounted for the overwhelming proportion of nonwhites in the South between 1940 and 1960. The participation rate for 1940 was obtained from U.S. Bureau of the Census (1943a p. 153). The state volumes of the 1950–1970 censuses are sufficiently detailed to calculate participation rates by enrollment status and, for most Southern states, within the urban, rural nonfarm, and farm sectors, which can then be aggregated (U.S. Bureau of the Census, 1952b table 71; 1963 table 117; 1973b table 166).

The South consists of 16 states (Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia) plus the District of Columbia.[24]

State-Level Minimum-Wage Coverage Rates in 1950.—For the census week in 1950 we estimated the fraction of employed black teenage males, aged 16–19, in each Southern state who were covered by the minimum-wage provisions of the Fair Labor Standards Act. The published volumes of the 1950 census present data on the occupation distribution of nonwhite males, classified by age. The occupational breakdowns consist of 180 categories, including major industry subsets of operatives and laborers. A large share of black teenage males were employed in the small number of categories for which industry-occupation coverage rates could be estimated from other sources.[25]

The first step in calculating the coverage rate was to use U.S. Bureau of the Census (1956a) to identify each occupational group in the state volumes that contained at least 0.2 percent of all employed nonwhite teenage males in 1950. Second, for occupational groups that were entirely (or almost entirely) employed in a single industry (e.g., sawyers, plasterers) we assigned the coverage ratio from U.S. Department of Labor (1948). By far the biggest employer of black male teens in 1950 was agriculture, for which the coverage rate in 1950 was zero.

Third, for occupational groups that were employed in a number of industries but for which no industrial breakdown was given in the state volumes (e.g., bakers, truck drivers, janitors) we used U.S. Bureau of the Census (1956b) to estimate the industry–occupation matrix.[26] Because we know the coverage rates for each industry (and we have allocated occupations to industry) we can estimate covered employment by occupation. The overall coverage rate is the sum of covered employment across occupations divided by the number of employed black teens.

Table 3.—All data are from U.S. Bureau of the Census (1952b).

REFERENCES

Alston, Lee, *Costs of Contracting and the Decline of Tenancy in the South*, New York: Garland, 1985.

Bond, Horace Mann, *The Education of the Negro in the American Social Order*, New York: Prentice-Hall, 1934.

Cogan, John, "The Decline in Black Teenage Employment, 1950–1970," *American Economic Review*, September 1982, *72*,

[23]Because the number of nonfarm Southern black teens in the 1900 census sample was very small, their participation rates by enrollment status and the enrollment rate (see Table 1) were calculated by combining the 1900 and 1910 public-use samples (Population Studies Center, 1989).

[24]Delaware is excluded from all analyses of 1950 data because no state-level data on nonwhites were reported.

[25]These sources are: (i) table 14 of U.S. Department of Labor (1948), which shows total nonsupervisory employment and the number of employees subject to the minimum-wage provisions of the Fair Labor Standards Act by industry group in September 1947; (ii) Charles H. Livengood (1951), which identifies the categories of employees who were exempt from the minimum-wage provisions as of 1950; (iii) table 7 of U.S. Bureau of the Census (1956a), which cross-classifies employed nonwhite males in the United States by age and detailed occupation (in 1950 more than 85 percent of all employed black teens lived in the South); (iv) table 2 of U.S. Bureau of the Census (1956b), which provides a cross-classification of all employed males (all ages and races combined) by detailed occupation and a partial breakdown of industries.

[26]The industrial classifications in U.S. Bureau of the Census (1956b) refer to all employed males. Because black teens are likely to be underrepresented relative to all employed males in highly covered industries (e.g., durable-goods manufacturing) the coverage rates are biased upward.

621–38.

Day, Richard, "The Economics of Technological Change and the Demise of the Sharecropper," *American Economic Review*, June 1967, *57*, 427–49.

Durand, John, *The Labor Force in the United States, 1890–1960*, New York: Social Science Research Council, 1948.

Lebergott, Stanley, *Manpower in Economic Growth: The American Record Since 1800*, New York: McGraw-Hill, 1964.

Livengood, Charles H., *The Federal Wage and Hour Law, Reflecting the Fair Labor Standards Amendments of 1949*, Philadelphia, PA: American Law Institute, 1951.

Mare, Robert D. and Winship, Christopher, "The Paradox of Lessening Racial Inequality and Joblessness Among Black Youth: Enrollment, Enlistment, and Employment, 1964–1981," *American Sociological Review*, February 1984, *49*, 39–55.

Margo, Robert A., *Race and Schooling in the South, 1880–1950: An Economic History*, Chicago: University of Chicago Press, 1990.

Musoke, Moses, "Mechanizing Cotton Production in the United States: The Tractor 1915–1960," *Explorations in Economic History*, October 1981, *18*, 347–75.

Osterman, Paul, *Getting Started: The Youth Labor Market*, Cambridge, MA: MIT Press, 1980.

Smith, James, "Race and Human Capital: Reply," *American Economic Review*, December 1986, *76*, 1225–9.

Welch, Finis, "Education and Racial Discrimination," in Orley Ashenfelter and Al Rees, eds., *Discrimination in Labor Markets*, Princeton, NJ: Princeton University Press, 1973, pp. 43–81.

Whatley, Warren, "Labor For the Picking: The New Deal in the South," *Journal of Economic History*, December 1983, *43*, 905–29.

_____, "Southern Agrarian Labor Contracts as Impediments to Cotton Mechanization," *Journal of Economic History*, March 1987, *47*, 45–70.

Wright, Gavin, *Old South, New South*, New York: Basic Books, 1986.

Center for Studies in Demography and Ecology, University of Washington, *United States Census Data, 1900: Public Use Sample*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 1980.

Federal Security Agency, U.S. Office of Education, *Biennial Surveys of Education in the United States, 1938–40 and 1940–42*, Washington, DC: U.S. Government Printing Office, 1947.

Population Studies Center, University of Pennsylvania, *Public Use Sample: 1910 Census of Population*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 1989.

U.S. Bureau of the Census, *Negro Population, 1790–1915*, Washington, DC: U.S. Government Printing Office, 1918.

_____, *Fourteenth Census of the United States Taken in the Year 1920, Volume II. Population 1920, General Report and Analytical Tables*, Washington, DC: U.S. Government Printing Office, 1922.

_____, *Fourteenth Census of the United States Taken in the Year 1920, Volume IV. Population 1920, Occupations*, Washington, DC: U.S. Government Printing Office, 1923.

_____, *Fifteenth Census of the United States 1930, Unemployment, Volume 1*, Washington, DC: U.S. Government Printing Office, 1931.

_____, *Fifteenth Census of the United States 1930, Population, Volume IV. Occupations, By States*, Washington, DC: U.S. Government Printing Office, 1933.

_____, *Negroes in the United States, 1920–1932*, Washington, DC: U.S. Government Printing Office, 1935.

_____, (1943a) *Sixteenth Census of the United States: 1940, Population, Volume IV. Characteristics by Age, Part 1, United States Summary*, Washington, DC: U.S. Government Printing Office, 1943.

_____, (1943b) *Sixteenth Census of the United States: 1940, Population, Volume IV. Characteristics by Age, Parts 2, 3, and 4*, Washington, DC: U.S. Government Printing Office, 1943.

_____, (1952a) *1950 Census of Agriculture, Volume II. General Report, Statistics by Subject*, Washington, DC: U.S. Government Printing Office, 1952.

_____, (1952b) *Census of Population: 1950, Volume II. Characteristics of the Population, Parts 2, 3, and 4*, Washington, DC: U.S. Government Printing Office, 1952.

_____, *Census of Population: 1950, Volume II. Characteristics of the Population, Part 1, United States Summary*, Washington, DC: U.S. Government Printing Office, 1953.

_____, (1956a) *United States Census of Population: 1950. Special Reports: Occupational Characteristics*, Washington, DC: U.S. Government Printing Office, 1956.

_____, (1956b) *United States Census of Population: 1950. Special Reports: Occupation By Industry*, Washington, DC: U.S. Government Printing Office, 1956.

_____, *Census of Population: 1960, Volume I. Characteristics of the Population, Part 2*, Washington, DC: U.S. Government Printing Office, 1963.

_____, (1973a) *1970 Census of Population, Volume 1. Characteristics of the Population, Part 1; United States Summary, Section 2*, Washington, DC: U.S. Government Printing Office, 1973.

_____, (1973b) *1970 Census of Population, Volume 1. Characteristics of the Population, Part 2*, Washington, DC: U.S. Government Printing Office, 1973.

U.S. Department of Health, Education, and Welfare, *Biennial Survey of Education in the United States, 1948–50*, Washington, DC: U.S. Government Printing Office, 1954.

U.S. Department of Interior, Office of Education, *Biennial Survey of Education in the United States, 1928–30*, Washington, DC: U.S. Government Printing Office, 1932.

U.S. Department of Labor, *Annual Report of the Wage and Hour and Public Contracts Division*, Washington, DC: U.S. Government Printing Office, 1948.

Causes of Intercity Variation in Homelessness

By MARJORIE HONIG AND RANDALL K. FILER*

Homelessness in America has become a major policy concern in recent years. Following estimates by a number of researchers in the 1980's that suggested as many as a half-million homeless,[1] the 1990 census places the current number at around a quarter of a million. Efforts to design policies to deal with this problem have been handicapped by a lack of systematic analysis of the causes of homelessness.[2] Policymakers have had little guidance from researchers in determining the relative importance of such potential causes of homelessness as tight housing markets, slack labor markets, reductions in real public-assistance benefits, tightening of eligibility requirements for public assistance, and noninstitutionalization of the mentally ill.

To assess the relative importance of these and other factors, we used estimates by the Department of Housing and Urban Development (HUD) of the homeless population in a cross-section of metropolitan areas in 1984. We do not attempt in this paper to develop a full structural model of homelessness. Rather, we present reduced-form findings as a first attempt to measure comprehensively and systematically the underlying causes of homelessness.[3]

We assume in this paper that homelessness results from an imbalance between the cost of available housing and a household's income. Such an imbalance may occur, for example, when housing markets are tight relative to labor markets and housing costs are therefore high relative to earnings (or to alternative resources, such as public assistance). Investigation of the causes of homelessness must go beyond housing markets alone, however, because of the special characteristics of the population at risk and the public policies that address their needs. Transfer payments and policies regarding institutionalization of the mentally ill, for example, should be important determinants of the incidence of homelessness but are not part of a standard housing model. Since homelessness represents the end of a spectrum of poor housing outcomes, we also estimate equations for two related conditions, "crowded" and "doubled-up" housing. These are often cited as causes of homelessness but are, in fact, different manifestations of the same underlying relationship between housing costs and household resources.[4]

*Department of Economics, Hunter College/CUNY, 695 Park Avenue, New York, NY 10021. This research was supported by the Institute for Research on Poverty under the Small Grants Program and the University Committee on Research of the City University of New York. The authors thank Steve Bartolomei-Hill, Howard Chernick, Steven Craig, Martha Hill, Charles Manski, Kathryn Nelson, and Cordelia Reimers for helpful comments, and Franco Pignataro for his unusually meticulous programming assistance.

[1]See Filer and Honig (1990) for a detailed discussion of estimates of the size and growth of the homeless population.

[2]Most studies have focused on the size of the homeless population and its demographic characteristics. A few analyses have examined a limited range of potential causes of homelessness. F. Stevens Redburn and Terry F. Buss (1986) examined the roles of population growth, climate, and housing conditions; William Tucker (1987) and John M. Quigley (1990) focused on the role of rent control. Karin Ringheim (1990) analyzed in depth a small sample of metropolitan areas.

[3]We view this effort as preliminary to a more detailed analysis we intend to undertake when complete data from the 1990 Census, which enumerates the homeless population, become available.

[4]Most previous economic analyses of housing markets have not focused on the decision of whether to share housing. Demographers, on the other hand, have long studied this issue when examining family formation and living arrangements. Among economists, a notable exception is the study by Axel Börsch-Supan and John Pitkin (1988), who focused on the choice between renting, purchasing, and sharing housing. We are concerned in this paper with the other margin, where individuals and families chose between sharing housing and entering the homeless shelter system. Such

I. Data

Although there has been considerable controversy concerning the reliability of the HUD data,[5] HUD's estimate of approximately 250,000–350,000 homeless individuals nationwide has been supported by several subsequent studies. These have used a variety of estimation strategies but generally have arrived at numbers in the range of the HUD estimates.[6] While HUD's local area estimates are likely to be imprecise, it is sufficient for our purposes that any errors are random across cities. We find no evidence that this is not the case. Furthermore, in almost every city where independent researchers conducted a full count of the homeless at the same time as the HUD survey, the enumeration fell within the most reliable range reported by HUD.[7] Our

TABLE 1—EXTENT OF POOR HOUSING OUTCOMES IN VARIOUS CITIES (RANKED ACCORDING TO EXTENT OF HOMELESSNESS)

City	Homelessness per 100,000 population	Crowding per 100,000 households	Doubling-up per 100,000 households
San Francisco, CA	535.1	4,300.6	1,259.9
Los Angeles, CA	412.0	10,106.6	1,803.1
Miami, FL	348.8	11,491.2	1,978.5
New York, NY	346.2	6,389.6	1,673.9
Chicago, IL	323.9	4,639.2	1,792.5
Worcester, MA	259.9	1,735.4	938.8
Fort Wayne, IN	208.1	2,051.7	1,520.4
Las Vegas, NV	205.0	4,383.8	1,688.7
Houston, TX	200.7	6,265.5	1,803.5
Seattle, WA	187.6	1,833.9	1,011.7
Detroit, MI	173.8	3,043.8	1,997.1
Reno, NV	147.7	2,901.8	1,165.7
Richmond, VA	147.6	2,224.4	1,690.5
Portland, OR	136.4	1,862.0	866.5
Hartford, CT	112.2	2,556.2	1,455.4
Little Rock, AR	103.5	3,744.9	1,892.2
Davenport, IA	94.5	1,982.0	1,031.3
Boston, MA	86.6	2,429.8	1,159.5
Tampa, FL	86.3	2,566.3	1,202.2
Philadelphia, PA	75.5	2,499.7	2,287.6
Lincoln, NE	65.3	830.8	451.0
Birmingham, AL	64.2	4,195.1	1,733.3
Phoenix, AZ	62.7	4,609.0	1,730.6
Cincinnati, OH	62.1	3,227.9	1,975.2
Columbia, MO	61.3	1,652.5	381.4
Louisville, KY	59.7	2,903.2	2,055.2
Danville, VA	58.5	3,784.6	1,649.7
Syracuse, NY	57.7	1,722.6	1,226.6
Grand Rapids, MI	55.9	1,857.9	947.0
Salt Lake City, UT	52.4	4,388.1	1,105.3
Sioux City, IA	46.5	2,323.2	454.6
Monroe, LA	45.6	5,136.1	1,571.5
Minneapolis/ St. Paul, MN	45.3	1,623.8	527.7
Raleigh/Durham, NC	42.2	2,855.3	1,304.7
Pittsburgh, PA	40.8	1,964.8	1,539.2
Pueblo, CO	38.0	3,258.1	1,712.6
Jackson, MI	32.7	1,769.9	919.0
Baton Rouge, LA	32.5	5,106.5	1,262.2
Dayton, OH	31.7	1,770.2	1,442.6
Athens, GA	31.1	2,874.0	847.5
Baltimore, MD	30.7	2,853.2	2,197.8
Tyler, TX	29.4	4,410.6	1,731.2
Colorado Springs, CO	27.9	2,090.4	939.9
Charlotte, NC	26.7	3,625.3	1,858.7
Kansas City, MO	25.1	2,092.8	1,446.2
Cleveland, OH	22.0	1,791.2	1,483.0
Binghamton, NY	19.9	1,305.8	1,282.1
Charleston, SC	17.5	3,937.5	1,778.2
Rochester, NY	13.6	1,255.1	1,116.5
Fall River, MA	6.8	3,125.0	1,250.0

a model was estimated using individual-level data by Renya Reed (1992). She suggests that most families prefer shared housing and enter the current homeless shelter system when constrained from this option. These questions can be further developed once the micro-level data from the 1990 Census are available.

[5]See Redburn and Buss (1986) for a useful summary of the methodology and estimation issues, and see Richard P. Applebaum (1987) and Lynn Parsons (1986) for methodological critiques.

[6]HUD used a mid-range based on four independent methods of arriving at a national estimate of between 250,000 and 350,000 homeless individuals in 1984. These methods included estimates from local studies, 500 key informant interviews in 60 metropolitan areas, surveys of 184 shelter operators in 60 metropolitan areas, and estimates of ratios of shelter and street populations (U.S. Department of Housing and Urban Development, 1984). An Urban Institute survey used a probability sample of service-using homeless individuals and estimated a total of 567,000–600,000 in 1987 (Martha R. Burt and Barbara E. Cohen, 1988; Burt, 1988). A study by ICF, Inc., based on the HUD data but differentiating between the metropolitan, suburban, and rural incidence of homelessness, estimated a homeless population of 355,000 in 1984 (Committee for Food and Shelter, 1987). Richard B. Freeman and Brian Hall (1987) derived an estimate of 279,000 in 1983, based on a New York City sample of the shelter population and estimates of the shelter-to-street ratio.

[7]See, for example, the surveys by Frederic G. Robinson (1985) for Washington, DC, Mark LaGory et al. (1989) for Birmingham, AL, and Franklin J. James (1989) for Denver. The one exception to this generalization is Peter H. Rossi (1989), who found considerably fewer homeless in Chicago than the HUD estimate.

measures of crowding and doubling-up pertain to all households in the 5-percent public-use sample of the 1980 Census. We use the standard definitions of "crowded" (more than one person per room) and "doubled-up" (households containing more than one nuclear family, defined as parents and their children related by blood or adoption) (Michael A. Stegman, 1988).

Table 1 indicates the numbers of homeless individuals (per 100,000 population) as well as the numbers of crowded and doubled-up households (per 100,000 households) for 50 of the 60 metropolitan areas surveyed by HUD.[8] The incidence of homelessness in these cities ranges from a high of 535.1 per 100,000 in San Francisco to 6.8 in Fall River, MA. Unfortunately, the HUD data do not distinguish between the two major components of the homeless population: single males and members of families headed by females. There is evidence from surveys of homeless populations in various metropolitan areas that the proportion of families varies considerably across cities (U.S. Conference of Mayors, 1987), but we are unable to capture these differences in this study. We include, however, independent variables that might be related to either or both of these two types of homelessness.

II. Results

Table 2 presents ordinary least-squares estimations of the incidence of homelessness, crowding, and doubling-up (per 100,000 population or households). The final column contains the population-weighted means and standard deviations of each variable.

We experimented with a number of measures to capture the characteristics of the low-rent housing market.[9] Rents were measured at the mean and at various points in the distribution for all apartments and for apartments of various sizes. We also used vacancy rates that corresponded to each rent level, as well as the fraction of housing in rental units, the growth and age of the housing stock, and the presence of rent control.[10] Among these, the level of rents at the tenth percentile of the rental distribution for all apartments had the greatest impact on the incidence of homelessness, higher rents being associated with higher rates of homelessness. An increase of one standard deviation in rents at the tenth percentile resulted in a predicted increase in homelessness of 78 persons per 100,000 population, or about 42 percent of the mean level of homelessness across the metropolitan areas in our sample. Since housing is a highly competitive market where rents should reflect average costs, this suggests that public policies to reduce the cost of providing minimally adequate housing could have a powerful impact in reducing homelessness. Among the obvious candidates for examination are building and zoning codes, tax assessment policies, and capital costs. Alternatively, increased rent subsidies could have the same impact, although with obvious implications for tax burdens.

Rents at the tenth percentile also affected the extent of crowding, but they were not significantly related to doubling-up. The corresponding vacancy rate (the fraction of apartments at or below the tenth percentile currently available for rent) appeared to be related to homelessness, although the estimated coefficient is not statistically significant at conventional levels. Vacancy rates

[8]We have combined Raleigh and Durham, NC, into a single observation to be consistent with the Census treatment of these two cities, and we have been forced to exclude Scranton, PA, Annapolis, MD, Bowling Green, KY, Hazelton, PA, Lewiston, ME, Lompoc, CA, Merced, CA, and Monroe, MI, because data on some or all of the independent variables were not available. Washington, DC, was also excluded because statewide variables do not exist.

[9]We assume that the alternative for individuals or families at risk of being homeless, or for those leaving crowded or doubled-up housing, is to rent rather than to purchasing housing.

[10]Most of the measures used are for the metropolitan area (SMSA) that contains the city for which HUD obtained data. Our ability to predict homelessness may be hampered by any cross-city variation in the relationship between the central city and the metropolitan area. We assume that labor and housing markets are metropolitan in scope but, for reasons of convenience and services, most homeless will locate in the core city of a metropolitan area.

TABLE 2—ESTIMATED CAUSES OF HOMELESSNESS, CROWDING, AND DOUBLING-UP

Independent variable	Dependent variable			
	Homeless	Crowded	Doubled-up	Mean [SD]
Rents at 10th percentile of all apartments	2.87 (3.93)	53.40 (4.33)	3.07 (1.19)	135.30 [27.26]
Vacancy rate at 10th percentile of all apartments	−872.90 (1.58)	952.58 (0.10)	3,141.06 (1.62)	0.059 [0.024]
Presence of rent-control law	−15.50 (0.23)	−224.78 (0.20)	−297.56 (1.26)	0.306 [0.461]
Growth in employment, 1980–1982	−859.09 (2.71)	2,910.76 (0.55)	−1,302.90 (1.17)	−0.016 [0.051]
Share of employment in service industries	−347.69 (1.33)	−9,365.56 (2.12)	1,272.58 (1.38)	0.272 [0.044]
Predicted size of low-skill labor market	−1,003.87 (0.38)	87,087.77 (1.98)	−775.71 (0.08)	0.417 [0.007]
Households below poverty line per 100,000 households	0.013 (1.22)	0.72 (4.03)	0.024 (0.64)	11,074.50 [2,287.28]
Local government expenditures on public welfare per capita	0.11 (0.59)	−5.21 (1.49)	−1.12 (1.54)	98.00 [136.71]
Maximum AFDC benefits, family of three	−0.95 (2.58)	−16.39 (2.64)	0.74 (0.57)	518.44 [86.22]
Maximum SSI benefits	1.07 (2.14)	24.00 (2.83)	3.60 (2.04)	359.98 [56.64]
Percentage reduction in AFDC benefits if living with nonpoor parents	146.62 (1.49)	1,628.93 (0.98)	−312.25 (0.90)	0.208 [0.242]
AFDC accuracy rate, 1983	98.15 (0.13)	40,168.43 (3.19)	−2,859.33 (1.09)	0.932 [0.025]
Mental health in-patients per 100,000 state population	−0.83 (1.50)	−0.03 (0.003)	2.63 (1.35)	102.37 [40.71]
Fraction of births to teenage mothers	−1,173.00 (1.39)	−14,645.25 (1.03)	9,633.92 (3.20)	0.128 [0.023]
Blacks per 100,000 population	0.004 (1.78)	−0.038 (1.01)	0.024 (3.03)	16,551.60 [8,212.83]
1984 population (100,000's)	1.22 (1.44)	49.90 (3.50)	2.54 (0.85)	38.14 [28.11]
Constant	307.54	−81,311.37	−25.25	
Adjusted R^2:	0.83	0.87	0.73	

Note: Numbers in parentheses are t statistics.

did not seem to affect the incidence of crowding: moreover, they were positively correlated with doubling-up.[11] Rent control, which has been cited as a cause of homelessness (Tucker, 1987),[12] had no effect on either homelessness or crowding but may have been associated with decreased doubling-up, although the effect is imprecisely estimated. Rents and vacancy rates at other points in the distribution and for apartments of various sizes had similar but less significant impacts on the housing outcomes. Other than these measures, no characteristic of housing markets had any observable effect on any of the outcome measures.

Turning to local labor markets, recent growth in private-sector employment showed a strong negative relationship to homelessness. It may also have reduced the incidence of doubling-up, although here the relationship is not precisely estimated. The relative size of the service sector was less strongly related to homelessness but significantly reduced crowding. In contrast, there was a positive, although not significant, relationship between this aspect of the local labor market and the extent of doubling-up.[13] The relative demand for low-skill labor[14] had little impact on homelessness or doubling-up but was positively associated with crowding rates. This suggests that in areas where a large fraction of the available jobs require low skill levels (and therefore presumably pay low wages), there is resulting pressure on the ability of workers to afford more than minimal-quality housing. In combination with the negative relationship between employment growth and homelessness, the absence of a relationship between the relative skill levels of an area's jobs and homelessness may suggest that even low-skill jobs provide sufficient income to avoid homelessness. This pattern of results suggests that local areas concerned with reducing homelessness may want to adopt strong pro-employment growth policies without regard to the type of employer being attracted.

Interestingly, the unemployment rate (overall or specific to the low-skill population), the long-term unemployment rate (unemployment of more than three months), and the ratio of employment to population, which has been found to be an important factor in youth labor markets (Freeman, 1982), did not have observable impacts on homelessness.

Although there was some suggestion that household incomes, as measured by the proportion of households in the metropolitan area with incomes below the poverty line, were related to homelessness, incomes were a much stronger predictor of crowding. However, there did not appear to be any relationship between incomes and the incidence of doubling-up. This finding is consistent with shared housing resulting from numerous personal and other factors in addition to those that create homelessness.

Government expenditures on social services would be expected to be related to conditions at the lower end of the housing market, and we experimented with a number of alternative measures.[15] While the

[11] Unlike crowding, doubling-up occurs throughout the housing market. The two functions differ considerably as a consequence. Nevertheless, we report results for doubling-up, since this phenomenon is often claimed to be related to homelessness.

[12] Our results for this variable are similar to those found by Quigley (1990) when he added price and income measures to Tucker's analysis of homelessness.

[13] In addition to low-skill service industries, this sector includes such high-wage industries as legal, medical (other than hospitals), and business services.

[14] This variable is defined as the predicted skill distribution in the metropolitan area. Specifically, it is the fraction of an area's workers who would have less than a high-school education if workers in each occupation in the area had the educational distribution that prevailed for those occupations in the nation as a whole. The actual skill composition of the labor market may be a flawed measure of the demand for low-skilled labor, since the observed fraction of workers with less than a high-school degree is likely to be strongly influenced by the pool of workers available in a local labor market.

[15] Both poverty rates and government expenditures on public welfare are likely to be at least partially endogenous, determined as a function of other included factors such as labor-market conditions and program benefit levels.

overall level of expenditures by localities on public welfare reduced the incidence of both crowding and doubling-up (although the coefficients were not significant at the 10-percent level), only program-specific benefits had any impact on homelessness.[16] Higher maximum Aid to Families with Dependent Children (AFDC) benefits were associated with lower rates of homelessness (as well as with a lower incidence of crowding). Supplemental Security Income (SSI) benefits, on the other hand, were associated with higher rates of not only homelessness but also crowding and doubling-up. These results are puzzling. It is possible that they may be explained by differences in the populations served by these two programs. Certainly the relationship between the SSI program and homelessness presents a fruitful area for future research.

We also included some program characteristics as proxies for the administrative stringency of public-assistance programs.[17] For example, the larger the statewide reduction in AFDC benefits for single mothers residing with nonpoor parents, the higher was the incidence of homelessness in the locality, although the coefficient was not statistically significant at the 10-percent level. Crowding and doubling-up, however, were, at most, weakly affected by this aspect of the AFDC program. Another measure, the statewide proportion of AFDC families that, according to a Department of Health and Human Services audit of state programs, were actually eligible for the benefits they were receiving (the AFDC accuracy rate) did not seem to influence homelessness. However, higher accuracy rates, which imply that fewer families were receiving AFDC benefits for which they were not eligible, were associated with higher rates of crowding.[18]

Treatment of the mentally ill, as measured by the number of inpatients in state mental-health facilities,[19] appears to have had an impact on the incidence of homelessness, although this effect is imprecisely measured. There has been heated controversy regarding the effect on homelessness of the decision in the 1960's to attempt treatment of the mentally ill in community centers rather than in state mental hospitals.[20] These findings suggest that the policy of not institutionalizing the mentally ill has been an important factor in increasing homelessness. While the incidence of crowding does not appear to be affected by hospitalization rates, we observe a positive (but again imprecisely measured) relationship with the extent of doubling-up.

Since we are unable to control for the number of mentally ill in each state, these coefficients require interpretation. The mentally ill individual faces three possible living situations: in a state hospital, on the street (homelessness), or being taken in by relatives (doubling-up). If the rate of mental illness were constant across metropolitan areas, the coefficients on hospitalization rates in the homeless and doubling-up equations should sum to -1 and would represent variations in how the mentally ill were allocated across living situations.

[16]Homeless families are likely to be eligible for AFDC benefits; homeless single individuals may be eligible for Social Security Disability Insurance (SSDI), a Federal program requiring previous employment, or SSI (a means-tested, combined federal and state program). SSDI benefits were not included, since benefit levels do not vary by locality.

[17]There is evidence for New York City, for example, that the majority of homeless families have at some time been AFDC recipients and that many had been dropped from the program for "administrative" reasons (e.g., for missing appointments or for failing to adhere to administrative requirements) (Anna Lou Dehavenon, 1989; James R. Knickman and Beth C. Weitzman, 1989).

[18]An additional measure, the number of SSI and SSDI recipients dropped from the rolls in the early 1980's, which has often been cited as a cause of homelessness, was not significant in any of the equations.

[19]Facilities are restricted to hospitals and do not include community care establishments. This measure was more significant than per capita state spending on mental health or a flow measure of the number of new admissions to mental-health facilities.

[20]See the discussions in W. Robert Curtis (1986) and Redburn and Buss (1986). In 1955, there were 559,000 patients in state mental hospitals; currently, there are approximately 123,000 (John A. Talbott, 1989).

Conversely, if hospitalization rates were constant but the rate of mental illness varied across areas, there would be positive coefficients on the number of hospitalized patients in equations predicting the extent of homelessness and doubling-up. The magnitude of these coefficients would be equal to the ratios of homeless or doubled-up individuals among the mentally ill to those hospitalized. Since it is likely that both mental illness and hospitalization rates vary across states, the observed coefficients combine the impact of variations in both rates. This means that the negative coefficient on the hospitalization rate in the homelessness equation understates the true effect of hospitalization policies by any positive influence from variations in the extent of mental illness. The positive relationship between hospitalization and doubling-up suggests that many more mentally ill individuals are cared for by friends and relatives than by state institutions.

We also included population characteristics that might affect the low-rent housing market even after controlling for housing markets, job opportunities, and incomes. The percentage of births to teenage mothers, for example, substantially increased the incidence of doubling-up but, surprisingly, appears to be related to lower levels of homelessness. The larger the relative size of the black population, the higher were the rates of homelessness and, especially, doubling-up.[21] Finally, the incidence of crowded housing, and possibly homelessness, was greater in larger metropolitan areas.[22]

[21]The remaining ethnic groups, Hispanics and Asians, are excluded because their density is heavily influenced by locational decisions of recent immigrants. These decisions are a function of labor markets, housing markets, and government policy, which are already in our equations. The exogenous variables included in the model explain 92 percent of the variation in the relative size of the Asian population in metropolitan areas, for example.

[22]A final factor, climate, which has often been cited as a cause of homelessness, was unrelated to homelessness.

III. Conclusion

Our key finding is that, despite the perceived weakness of the HUD data and the large variation in the incidence of homelessness and other poor housing outcomes across metropolitan areas,[23] a relatively parsimonious and intuitively appealing set of factors is able to explain variations in homelessness, crowding, and doubling-up across American cities. The findings indicate that homelessness has multiple causes. It has roots in housing markets, labor markets, and public policies regarding the treatment of the mentally ill and the low-income population. Although replication of this study using data from the 1990 Census should be fruitful, the results reported above are suggestive and provide a framework for discussions of policies to reduce homelessness.

REFERENCES

Applebaum, Richard P., *Counting the Homeless*, Chicago: Center for Urban Research and Policy Studies, University of Chicago, February 1987.

Börsch-Supan, Axel and Pitkin, John, "On Discrete Choice Models of Housing Demand," *Journal of Urban Economics*, September 1988, *24*, 153–72.

Burt, Martha R., "Comments," *Rethinking Policy on Homelessness*, Washington, DC: Heritage Foundation, 1988.

_____ **and Cohen, Barbara E.,** *Feeding the Homeless: Does the Prepared Meals Provision Help?* Vols. I and II, Washington, DC: Urban Institute, 1988.

Curtis, W. Robert, "The Deinstitutionalization Story," *Public Interest*, Fall 1986, *85*, 34–49.

Dehavenon, Anna Lou, "The Tyranny of In-

[23]The mean incidence of homelessness in the 1984 HUD data was 187.5 per 100,000 population with a standard deviation of 147.8. On average, 4,422.5 out of every 100,000 households in the cities in our sample were crowded (standard deviation of 2,762.2), and 1,617.3 were doubled-up (standard deviation of 409.8).

difference: A Study of Hunger, Homelessness, Poor Health and Family Dismemberment in 818 New York City Households with Children in 1988–89," mimeo, East Harlem Interfaith Welfare Committee, New York, 1989.

Filer, Randall K. and Honig, Marjorie, *Policy Issues in Homelessness: Current Understanding and Directions for Research*, New York: Manhattan Institute for Policy Research, 1990.

Freeman, Richard B., "Economic Determinants of Geographical and Individual Variations in the Labor Market Position of Young Persons," in Richard B. Freeman and David A. Wise, eds., *The Youth Labor Market Problem: Its Nature, Causes, and Consequences*, Chicago: University of Chicago Press, 1982, pp. 115–54.

_____ **and Hall, Brian,** "Permanent Homelessness in America?" *Population Research and Policy Review*, January 1987, *6*, 3–27.

James, Franklin J., "Factors Which Shape the Risks of Homelessness: Preliminary Observations From Colorado," mimeo, Denver: University of Colorado Graduate School of Public Affairs, September 1989.

Knickman, James R. and Weitzman, Beth C., *A Study of Homeless Families in New York City: Risk Assessment Models and Strategies for Prevention*, New York: Health Research Program of New York University, 1989.

LaGory, Mark, Ritchey, Ferris J., O'Donoghue, Timothy and Mullis, Jeffrey, "Homeless People in Alabama: A Variety of People and Experiences," in J. Momeni, ed., *Homelessness in The United States*, Westport, CT: Greenwood, 1989.

Parsons, L., *Literature Review of Studies Which Deal with Estimating the Size of the Homeless Population*, Amherst, MA: Social and Demographic Research Institute, University of Massachusetts, 1986.

Quigley, John M., "Does Rent Control Cause Homelessness? Taking the Claim Seriously," *Journal of Policy Analysis and Management*, Winter 1990, *9*, 89–93.

Redburn, F. Stevens and Buss, Terry F., *Responding to America's Homeless*, New York: Praeger, 1986.

Reed, Renya, "The Determinants of Homelessness Amongst Families on Welfare: Estimation of Three Econometric Models of Search in the Low Income Housing Market," Ph.D. dissertation, Johns Hopkins University, 1992.

Ringheim, Karin, *At Risk of Homelessness: The Roles of Income and Rent*, New York: Praeger, 1990.

Robinson, Frederic G., "Homeless People in the Nation's Capital," mimeo, University of the District of Columbia, 1985.

Rossi, Peter H., *Down and Out in America: The Origins of Homelessness*, Chicago: University of Chicago Press, 1989.

Stegman, Michael A., *Housing and Vacancy Report: New York City, 1987*, City of New York: Department of Housing Preservation and Development, April 1988.

Talbott, John A., "Testimony before U.S. Senate Subcommittee on Housing and Urban Affairs," *Access* (Washington, DC: National Resource Center on Homelessness and Mental Illness), 1989, *1*, 3.

Tucker, William, "Where Do the Homeless Come From?" *National Review*, 25 September 1987, *39*, 32–43.

Committee for Food and Shelter, *Characteristics and Housing Needs of the Homeless*, Washington, DC: ICF, Inc., 1987.

U.S. Conference of Mayors, *The Continuing Growth of Hunger, Homelessness and Poverty in America's Cities: 1987*, Washington, DC: U.S. Conference of Mayors, January 1987.

U.S. Department of Housing and Urban Development, *A Report to the Secretary on the Homeless and Emergency Shelters*, Washington, DC: Office of Policy Development and Research, May 1984.

Cyclical Consumption Patterns and Rational Addiction

By ENGELBERT J. DOCKNER AND GUSTAV FEICHTINGER*

Much of empirically observable consumption behavior seems to contradict rational-choice theory. Here we refer to heavy eating followed by strict dieting; smoking, quitting and starting again; taking drugs, quitting, and starting all over again; and so on. In a recent paper, Gary S. Becker and Kevin M. Murphy (1988) demonstrate that a wide variety of consumer behavior—addiction included—is consistent with utility maximization. As a consequence, addictive behavior need not be excluded from the rational-choice framework. As Becker and Murphy (1988) point out, an even stronger claim holds true: analyzing addictive behavior within the rational-choice framework provides new insight and a better economic understanding of addiction.

Explaining addictive behavior in terms of utility-maximizing consumers does not imply, however, that addicts can be considered as *happy* persons. What this framework does is to demonstrate that for a variety of reasons certain consumers can become addicted; once they are addicted, to stop consuming the addictive good would make them worse-off.

A consumer is said to be addicted to a consumption good c if an increase in past consumption of c causes present consumption to rise. In the theory of rational addiction (see George J. Stigler and Becker, 1977; Laurence R. Iannaccone, 1986; Becker and Murphy, 1988; Daniel Leonard, 1989) past consumption is summarized by a stock of *consumption capital* (habits) that, together with current consumption, affects current utility. Hence, a consumer can be called addicted to c if consumption of c increases with an increase in the stock of its corresponding consumption capital. This definition implicitly assumes that c accumulates a single stock (consumption capital). We call this stock *commodity-specific consumption capital*. It is this case that is extensively discussed in the literature on rational addiction. If, on the contrary, a consumption good accumulates several stocks of consumption capital, past consumption affects current utility through several channels. As one might expect, this requires a separate analysis of addiction. It is at this point that the present note steps in. We study a model of addiction in which a single consumption good accumulates two stocks. This is the simplest case without commodity-specific consumption capital, but it provides all the economic intuition for the more general case.

One important consequence of the relaxation of commodity-specific consumption capital is that the variety of possible consumption profiles over time is enlarged. This motivates us to explore the relationship between various degrees of addiction and the consumption profile over a person's lifetime. In the case of commodity-specific consumption capital, Becker and Murphy (1988) discuss several possible outcomes: they show, for example, that in the case of very strong addiction rational consumers have to go *cold turkey* (i.e., stop consuming the good abruptly). In other cases with lower degrees of addiction, consumption follows a monotonic path. This holds true even for the case

*Dockner: Department of Economics, Vienna University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria; Feichtinger: Institute for Econometrics, OR, and Systems Theory, Technical University of Vienna, Argentinierstr. 8, A-1040 Vienna, Austria. We thank Gary S. Becker, Franz Hof, Gerhard Sorger, Dan Usher, Franz Wirl, seminar participants at the Institute for Advanced Studies in Vienna and Queen's University, and three anonymous referees for their helpful comments. Research assistance was provided by Andreas Novak and Doris Ostrusska. The first author acknowledges financial support from Queen's University at Kingston, Canada, during his visit in 1990/1991 while this paper was being revised. The research of the second author was supported by a grant from the Austrian Science Foundation (no. P6601).

of an unstable steady state. However, does this imply that cyclical consumption policies cannot be explained by the theory of rational addiction? Certainly not, as has already been demonstrated by Becker and Murphy (1988). Our note elaborates on their analysis, and we fully explain the causes for the occurrence of cyclical consumption trajectories. We show that, for the case of commodity-specific consumption capital, consumption trajectories will always be monotonic. Hence, cyclical consumption paths expressed as stable limit cycles or damped (explosive) oscillations require a consumption good that accumulates at least two stocks. It is the interaction of these two stocks that causes *irregular* behavior. In particular, we show that only if present consumption is positively correlated with past consumption as expressed in one of the two stocks (say, a stock called *eating capital*) but is negatively correlated with the other stock (say, one called *weight*) cyclical consumption patterns are possible. This implies that consumption cycles require two counterbalancing effects: an addictive one and a satiating one. The addictive force causes current consumption to increase as past consumption accumulates (hence, the ascending part of the cycle); the satiating force causes current consumption to decline as habits accumulate (hence, the descending part of the cycle).

As mentioned above, cyclical consumption profiles can be expressed both in terms of damped or explosive oscillations as well as in terms of stable limit cycles. In the first case, the trajectory either converges to or diverges from the steady-state consumption capitals (i.e., the long-run solutions). In the latter case the long-run equilibrium is not a single point, but rather an invariant manifold (i.e., the limit cycle). Economically, the existence of a stable limit cycle implies that *binges* can continue to cycle throughout much of a person's lifetime. This brings us back to the examples discussed at the very beginning of this paper and explains why certain consumers might follow binges throughout much of their lifetimes.

In Section I, we introduce a model of rational addiction that is a variation of the two-state-variable example of Becker and Murphy (1988). Here we concentrate only on what might be called the *semi-reduced-form* model that involves the allocation of consumption goods over time and the corresponding accumulation of two stocks of consumption capital. In Dockner and Feichtinger (1989) we demonstrate that this model can be derived from a more general one that includes the intertemporal choice of leisure time and goods allocation subject to a wealth constraint. In Section II, we discuss the relationships among complementarity over time, addiction, and satiation and use these results to give precise information about monotonic or cyclical consumption profiles. In Section III, we present a numerical example with strictly concave preferences that leads to a stable limit cycle as optimal policy. Section IV concludes this note.

I. The Model

We consider a representative consumer who at each instant of time t derives utility from consumption of goods $c_1(t)$ and $c_2(t)$ and stocks of consumption capital $S_1(t)$ and $S_2(t)$, as in

$$(1)\quad U(t) = U\big(c_1(t), c_2(t), S_1(t), S_2(t)\big).$$

We assume that the utility function $U(c_1, c_2, S_1, S_2)$ is strictly concave jointly in (c_1, c_2, S_1, S_2). Moreover, the cross partial derivative $U_{S_iS_2}$ is set equal to zero (i.e., marginal utility with respect to consumption capital S_i is independent of the level of S_j, $i \neq j$).

The stocks $(S_1(t), S_2(t))$ are measures of past consumption of $(c_1(t), c_2(t))$ that affect current utility through an accumulation process; that is,

$$(2)\quad \dot{S}_1(t) = f_1\big(c_1(t), c_2(t)\big) - \delta_1 S_1(t)$$

$$(3)\quad \dot{S}_2(t) = f_2\big(c_1(t), c_2(t)\big) - \delta_2 S_2(t).$$

Relations (2) and (3) are neoclassical accumulation equations with δ_i as constant depreciation rates, and the functions f_i are

specified depending on whether the consumption capitals are commodity-specific or not. Throughout the paper we assume that $\delta_1 > \delta_2$.

Assuming an infinitely-lived consumer with a constant rate of time preference, r, who maximizes the discounted stream of utility subject to the accumulation equations (2) and (3), the dynamic household problem becomes

$$(4)\quad \max_{c_1(t),c_2(t)} \int_0^\infty e^{-rt} U(c_1(t), c_2(t), S_1(t), S_2(t))\, dt$$

subject to (2) and (3) and given initial conditions $S_i(0) = S_{i0} \geq 0$.

Model (2)–(4) describes the optimal allocation of the consumption good(s) over time when utility depends on both current and past levels of consumption. It can be derived from a more general model that includes the intertemporal choice of leisure as well as a wealth constraint (see Dockner and Feichtinger [1989] for details).

The general specification (2)–(4) allows for two important examples as special cases: one with commodity-specific consumption capital and one without it. Adopting the definition of Iannaccone (1986), we call the consumption capitals (habits) *commodity-specific* if for each S_i there exists a unique commodity c_i, such that f_i and U_{c_i}/U_{S_i} depend only on c_i and S_i. In our case with two consumption capitals this implies that each consumption good accumulates a single stock, $f_i(c_i(t), c_j(t)) = c_i(t)$, and that the marginal rate of substitution between current and past consumption is a function of c_i and S_i. In order for the latter condition to be satisfied together with $U_{S_1S_2} = 0$, we restrict attention to the additively separable case $U = U_1 + U_2$. To sum up, the case of commodity-specific consumption capital is represented by the accumulation equations $\dot{S}_i(t) = c_i(t) - \delta_i S_i(t)$ and the utility function $U = U_1(c_1, S_1) + U_2(c_2, S_2)$.

In the case with no commodity-specific consumption capital we restrict attention to the simplest case possible and assume that a single good, $c(t)$, accumulates two distinct consumption capitals—eating capital (S_1) and weight (S_2)—and that preferences are of the general form represented by (1). Algebraically this amounts to

$$(5)\quad \dot{S}_1(t) = c(t) - \delta_1 S_1(t)$$

$$\dot{S}_2(t) = c(t) - \delta_2 S_2(t)$$

with preferences given by $U(c, S_1, S_2)$.

The optimal paths of $c(t)$ (the case of no commodity-specific consumption capital) or $(c_1(t), c_2(t))$ (commodity-specific consumption capital) and $(S_1(t), S_2(t))$ are determined by the first-order conditions

$$(6)\quad c \text{ or } (c_1, c_2) = \arg\max\left\{U + \lambda_1 \dot{S}_1 + \lambda_2 \dot{S}_2\right\}$$

$$(7)\quad \dot{\lambda}_1 = (r + \delta_1)\lambda_1 - U_{S_1}$$

$$(8)\quad \dot{\lambda}_2 = (r + \delta_2)\lambda_2 - U_{S_2}$$

and the transversality condition

$$(9)\quad \lim_{t\to\infty} e^{-rt}\left\{\lambda_1(t)\left[\tilde{S}_1(t) - S_1(t)\right] + \lambda_2(t)\left[\tilde{S}_2(t) - S_2(t)\right]\right\} = 0$$

for all feasible states $\tilde{S}_1(\cdot), \tilde{S}_2(\cdot)$. In the above equations, λ_i are the shadow prices of the consumption stocks S_i. They measure the value of the future benefits or costs of consumption. Through the price system (λ_1, λ_2), rational consumers take into account the future consequences of current actions. Depending on whether consumption capital has beneficial or harmful effects, we expect the prices to be positive or negative.

II. Addiction and Cyclical Consumption Paths

A commodity, c, is called *addictive* if its current consumption increases as habits derived from its previous consumption accu-

mulate. Hence, in the case of commodity-specific consumption capital a consumer displays addictive behavior if $c_i(t)$ increases with an increase in its corresponding $S_i(t)$. This occurs if preferences are characterized by *adjacent complementarity*.[1]

The case of non-commodity-specific habits requires a separate treatment. First, we have to present a formal definition of addiction. Second, we need to explore whether the relationship between addiction and adjacent complementarity carries over to this case. This is not a trivial issue because past consumption affects current utility and hence current consumption through several channels, and therefore the interaction of the stocks will be of importance. In our example it is the current level of both eating capital and weight that influence $c(t)$.

To shed some light on these open questions we concentrate on the optimal path for the simplest case with nonspecificity as summarized through (5). Introducing a transformation of variables such that the steady-state capital stocks and the steady-state level of the single consumption good are zero and allowing for linearizations, the dynamics of the optimal solution is governed by

$$(10) \qquad S_1(t) = a_{11}e^{\mu_1 t} + a_{12}e^{\mu_2 t}$$

$$(11) \qquad S_2(t) = a_{21}e^{\mu_1 t} + a_{22}e^{\mu_2 t}$$

where μ_1 and μ_2 are the stable roots (or the two smaller roots in the case of instability) of the linearized modified Hamiltonian system. These dynamics imply that, locally around the steady state, the optimal $c(t)$ is a linear function of both $S_1(t)$ and $S_2(t)$; that is,

$$(12) \quad c(t) = \alpha S_1(t) + \beta S_2(t)$$

$$= \frac{(\mu_1 + \delta_1)(\mu_2 + \delta_1)}{\delta_1 - \delta_2} S_1(t)$$

$$- \frac{(\mu_2 + \delta_2)(\mu_1 + \delta_2)}{\delta_1 - \delta_2} S_2(t)$$

(see the Appendix for derivations). This relationship allows for a general discussion of addictive and other behavior in the case of non-commodity-specific consumption capital. We call a consumer *fully addicted* to c if both stocks of consumption capital are positively correlated to c (i.e., if α and β are positive). In terms of our example, a consumer is fully addicted to c if both an increase in eating capital S_1 and an increase in weight S_2 cause him to consume more of c. This behavior will lead to a monotonic consumption profile and is consistent with convergence to the steady state (stability). A consumer need not be fully addicted to c, but instead may display addictive and satiating behavior at the same time. We call this behavior *partially addicted*. This occurs if one of the stocks is positively correlated with c and the other one is negatively correlated with c. As we will see, this behavior most likely results in consumption cycles. Before discussing cyclical consumption patterns, however, let us answer the question of what causes a consumer to become fully addicted if habits are not commodity-specific. The next proposition provides an answer to this question.

PROPOSITION 1: *If consumption capitals S_1 and S_2 are not commodity-specific, then a consumer is fully addicted to c (i.e., α and β are both positive) if and only if his behavior displays adjacent complementarity with respect to both stocks.*

(See the Appendix for a proof.)

[1]The relationship between addiction and adjacent complementarity in the case of commodity-specific consumption capital first appeared explicitly in Marcel Boyer (1983) and was discussed at length by Iannaccone (1986) and Becker and Murphy (1988). It appeared implicitly but was not explicitly recognized in Harl E. Ryder and Geoffrey M. Heal (1973).

Proposition 1 generalizes the relationship between addiction and adjacent complementarity to non-commodity-specific habits. It also reveals that addiction implies that past consumption as measured in S_1 and S_2 raises the marginal utility of current consumption (i.e., $U_{cS_i} > 0$).

The proof of Proposition 1 also shows that addiction with respect to both habits implies monotonic consumption paths and that it is consistent with convergence to the steady state: as long as we have $-\delta_1 < \mu_1 < -\delta_2$ and $0 > \mu_2 > -\delta_2$, the consumer displays addictive behavior that results in monotonic convergence to the steady state.

As is clear from equation (12) cyclical consumption paths require that one of the stocks is positively correlated with c and that the other one is negatively correlated with c. These cycles can be expressed as damped/explosive oscillations or limit cycles, depending on whether the eigenvalues μ_1 and μ_2 are complex with negative (damped oscillations), positive (explosive oscillations), or zero (limit cycles) real parts. As has been demonstrated by Becker and Murphy (1988) the roots μ_1 and μ_2 are conjugate complex if preferences display adjacent complementarity with respect to the stock with the higher depreciation rate and distant complementarity otherwise. In any case, conjugate complex roots imply that consumption is positively correlated with S_1, the stock with the higher depreciation rate, and negatively correlated with S_2. In light of equation (12), this allows for an interesting economic interpretation. Cyclical consumption patterns require two counterbalancing forces: an addictive force that causes current consumption to rise as past consumption accumulates and a satiating force that causes it to fall. The higher depreciation rate of the addictive stock then generates periods with increasing as well as decreasing consumption.

This discussion summarizes the existence of cyclical consumption paths in a model without commodity-specific consumption capital. What remains to be shown is whether consumption cycles are possible with commodity-specific habits.

PROPOSITION 2: *If consumption capital is commodity-specific, optimal consumption paths are monotonic.*

(See the Appendix for a proof.)

This is an interesting result and underlines the importance of nonspecificity for cyclical consumption profiles. Consumption cycles are possible only if a single consumption good accumulates two stocks of consumption capital, one of which generates addictive and the other one satiating behavior. These cycles, however, are the result of forward-looking behavior. Only a consumer with a desire to eat and a dislike for weight who anticipates the future consequences of his current actions can end up in eating and dieting cycles.

III. Persistent Consumption Cycles: A Numerical Example

The general discussion carried out in the preceding section has revealed that different configurations for optimal policies are possible depending on the nature of the consumption capital (whether it is commodity-specific or not) as well as on complementarity in consumption. In particular, we have seen that the theory of rational addiction is capable of explaining cyclical consumption paths expressed as damped/explosive waves or limit cycles. In this section we present a numerical example that establishes a stable limit cycle as optimal policy.[2] Our interest in limit cycles as optimal consumption paths is motivated by the fact that a *stable* limit cycle can be identified as persistent oscillatory behavior and hence is capable of explaining binges that continue to cycle throughout much of a person's lifetime.

We have to restrict our analysis to the case without commodity-specific consump-

[2]It should be noted that Marcel Boyer (1978) does mention the possibility of a limit cycle as an optimal consumption path but does not characterize it. Becker and Murphy (1988) characterize cyclical consumption in terms of damped/explosive waves but do not discuss limit cycles.

tion capital. We specify the utility function as

$$(13)\quad U(c,S_1,S_2)=c^{\eta}S_1^{\varepsilon}+\frac{\theta}{2}c^2+\frac{\nu}{2}S_1^2+\frac{\xi}{2}S_2^2+\pi cS_1.$$

The utility function (13) is strictly concave jointly in (c,S_1,S_2) as long as $0<\eta,\varepsilon<1$, $\eta+\varepsilon<1$, $\theta<0$, $\nu<0$, $\xi<0$, and $\theta\nu-\pi^2>0$. As the inequality for the marginal utility of consumption ($U_c>0$) is not globally satisfied for (13) we restrict our numerical calculations to the region of the (c,S_1,S_2)-space where it holds.

For the numerical example we make use of the following parameter values:

$$(14)\quad \delta_1=0.45\quad \delta_2=0.14\quad \eta=0.50$$
$$\varepsilon=0.49\quad \theta=-2\quad \nu=-16$$
$$\xi=-10\quad \pi=5.5.$$

To prove the existence and stability of the limit cycle we apply the Hopf bifurcation theorem (see John Guckenheimer and Philip Holmes [1990] for details). Among other things, this requires the existence of a pair of imaginary roots μ_1 and μ_2. As noted, pure imaginary roots result in a positive correlation between c and S_1 and a negative correlation between c and S_2.

PROPOSITION 3: *If the parameters are specified as in (14) and the time preference rate assumes the critical value* $r_0=5.8218875$, *then there exists a pair of imaginary roots that gives rise to the local existence of limit cycles. These cycles are stable and occur for an interval of time preference rates r slightly greater than* r_0.

PROOF:

A detailed numerical analysis of this result can be found in Dockner and Feichtinger (1989).

Proposition 3 provides us with the existence of stable limit cycles as rational intertemporal consumption choices. These cycles are *generic* in the sense that they do occur for an *interval* of time preference rates (parameter values) (see Guckenheimer and Holmes, 1990). Hence, limit cycles are *not knife-edge*, as perhaps suggested by the critical value r_0.

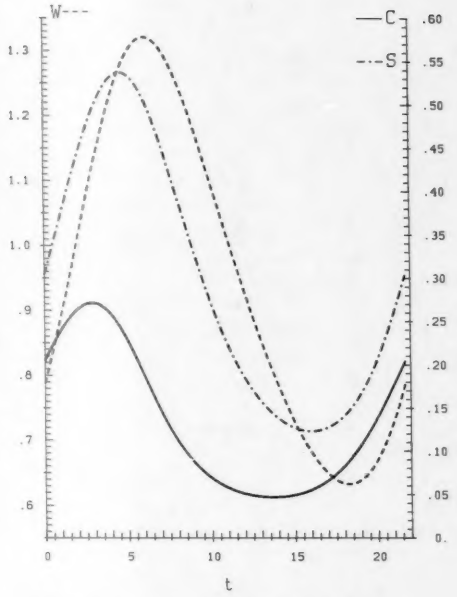


FIGURE 1. OPTIMAL CONSUMPTION (c), EATING CAPITAL (S), AND WEIGHT (W) PROFILES OVER TIME

The cyclical trajectories corresponding to the stable limit cycle are depicted in Figure 1. These paths show the following characteristics. Initially, both eating capital and weight are low. Given the positive correlation between S_1 and c, consumption rises. While c increases eating capital rises more rapidly than weight. Eventually, consumption levels off while weight continues to increase. With declining consumption and a high rate of depreciation, eating capital declines. Dieting also causes weight to decline so that, after some time, both eating capital and weight are at their initial values, and the cycle starts all over again.

IV. Concluding Remarks

This paper presents two new propositions in the theory of rational addiction. First, we show that the notion of commodity-specific consumption capital is crucial for the understanding of cyclical consumption paths. Only a model of rational addiction without commodity-specific consumption capital is capable of explaining consumption cycles. Second, we demonstrate that consumption behavior may very well end up in persistent oscillations (i.e., a limit cycle). We derive this result in terms of a numerical example and use it to explain binges that continue to cycle much throughout a person's lifetime.

Appendix

Derivation of Equation (12).—Differentiation of (10) and (11) with respect to time t and substituting equation (5) yields

$$c(t) = (a_{11}\mu_1 + a_{11}\delta_1)e^{\mu_1 t} + (a_{12}\mu_2 + a_{12}\delta_1)e^{\mu_2 t}$$

$$c(t) = (a_{21}\mu_1 + a_{21}\delta_2)e^{\mu_1 t} + (a_{22}\mu_2 + a_{22}\delta_2)e^{\mu_2 t}.$$

Substituting (10) and (11) into $c(t) = \alpha S_1(t) + \beta S_2(t)$ yields

$$c(t) = (\alpha a_{11} + \beta a_{21})e^{\mu_1 t} + (\alpha a_{12} + \beta a_{22})e^{\mu_2 t}.$$

Comparing coefficients in the above three equations results in

$$\alpha a_{11} + \beta a_{21} = a_{11}(\mu_1 + \delta_1)$$

$$\alpha a_{12} + \beta a_{22} = a_{12}(\mu_2 + \delta_1)$$

$$\alpha a_{11} + \beta a_{21} = a_{21}(\mu_1 + \delta_2)$$

$$\alpha a_{12} + \beta a_{22} = a_{22}(\mu_2 + \delta_2)$$

with a unique solution

$$\left(\alpha, \beta, \frac{a_{11}}{a_{21}}, \frac{a_{22}}{a_{12}}\right)$$

that provides us with the coefficients in equation (12).

PROOF OF PROPOSITION 1:

From Dockner and Feichtinger (1991), we know that the stable roots are given by

$$\mu_1 = \frac{r}{2} - \sqrt{\left(\frac{r}{2}\right)^2 - \frac{K}{2} + \frac{1}{2}\sqrt{K^2 - 4\det J}}$$

$$\mu_2 = \frac{r}{2} - \sqrt{\left(\frac{r}{2}\right)^2 - \frac{K}{2} - \frac{1}{2}\sqrt{K^2 - 4\det J}}$$

where

$$K = -\gamma_1 - \gamma_2 + A_1 + A_2$$

$$\det J = \gamma_1\gamma_2 - \gamma_1 A_2 - \gamma_2 A_1$$

with $\gamma_i \equiv \delta_i(\delta_i + r)$ and

$$A_i \equiv -\frac{1}{U_{cc}}\left[(r + 2\delta_i)U_{cS_i} + U_{S_iS_i}\right].$$

There is said to be *adjacent complementarity* with respect to stock S_i if $A_i > 0$ and *distant complementarity* if $A_i < 0$ (see Ryder and Heal, 1973).

According to equation (12) and keeping in mind that $\delta_1 > \delta_2$, α and β are both positive if and only if $\mu_1 + \delta_1 > 0$, $\mu_2 + \delta_2 > 0$, and $\mu_1 + \delta_2 < 0$ hold simultaneously. It is easily shown that this, however, is equivalent with $A_1 > 0$ and $A_2 > 0$.

PROOF OF PROPOSITION 2:

Iannaccone (1986) has shown that with commodity-specific consumption capital the optimal path of S_i near the steady state is approximated by $S_i(t) = S_i^* + k_i e^{\mu_i t}$ where S_i^* is the steady-state level, k_i is a constant, and μ_i is a stable eigenvalue given by

$$\mu_i = \frac{r}{2} - \sqrt{(r/2 + \delta_i)^2 - A_i}.$$

As is shown in Dockner and Feichtinger (1989) strict concavity of preferences implies $(r/2+\delta_i)^2 - A_i > 0$. Hence $S_i(t)$ can never be cyclical. Noting that, in the neighborhood of the steady state, $\text{sgn}(\dot{c}_i) = \text{sgn}(A_i)\text{sgn}(\dot{S}_i)$ implies that consumption can never be cyclical, too.

REFERENCES

Becker, Gary S. and Murphy, Kevin M., "A Theory of Rational Addiction," *Journal of Political Economy*, July 1988, *96*, 675–700.

Boyer, Marcel, "A Habit Forming Optimal Growth Model," *International Economic Review*, October 1978, *19*, 585–609.

_____, "Rational Demand and Expenditures Patterns under Habit Formation," *Journal of Economic Theory*, October 1983, *31*, 27–53.

Dockner, Engelbert J. and Feichtinger, Gustav, "Cyclical Consumption Patterns and Rational Addiction," Working Paper No. 120, Technical University of Vienna, May 1989.

_____ **and** _____, "On the Optimality of Limit Cycles in Dynamic Economic Systems," *Journal of Economics*, January 1991, *53*, 31–50.

Guckenheimer, John and Holmes, Philip, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, New York: Springer-Verlag, 1990.

Iannaccone, Laurence R., "Addiction and Satiation," *Economics Letters*, January 1986, *21* (1), 95–9.

Leonard, Daniel, "Market Behavior of Rational Addicts," *Journal of Economic Psychology*, January 1989, *10*, 117–44.

Ryder, Harl E. and Heal, Geoffrey, M., "Optimal Growth with Intertemporally Dependent Preferences," *Review of Economic Studies*, January 1973, *40*, 1–31.

Stigler, George J. and Becker, Gary S., "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, *67*, 79–90.

The Uncertain Unit Root in Real GNP

By GLENN D. RUDEBUSCH*

The impulse and propagation mechanisms of business cycles have long been debated; however, until recently, economists were in fairly broad agreement that business fluctuations could be studied separately from the secular growth of the economy. This separation was justified because, to a first approximation, the factors underlying trend growth were assumed to be stable at business-cycle frequencies. Indeed, the common practice of macroeconomists of all theoretical persuasions was to model movements in real GNP as stationary fluctuations around a linear deterministic trend (e.g., Finn Kydland and Edward C. Prescott, 1980; Olivier J. Blanchard, 1981). Such a trend-stationary (TS) model of real GNP was the canonical empirical representation of aggregate output until the early 1980's.

In contrast to previous work, much of the research of the last ten years has assumed a unit root in the autoregressive representation of real GNP, which is inconsistent with a TS model of output. A model with a unit root, commonly termed a "difference-stationary" (DS) model, implies that any stochastic shock to output contains an element that represents a permanent shift in the level of the series. If real GNP is best represented by a DS model, the traditional separation between business cycles and trend growth is incorrect. In the usual empirical versions of the DS model estimated for real GNP, output behaves more like a random walk than like transitory deviations from a steadily growing trend (e.g., John Y. Campbell and N. Gregory Mankiw, 1987; Peter K. Clark, 1987). The long-run trend of these DS representations is not fixed, as in a TS model, but stochastic; indeed, in the typical DS model, almost all fluctuations in output represent permanent shifts in trend rather than transitory movements in cycle. From another perspective, the essential difference between the two models can be found in the persistence of their dynamic responses to random shocks. In the DS model of output, the effect of a shock persists forever because the disturbance changes the trend component and thus affects the level of output in all future periods. In contrast, the impact of a shock in the TS model is transitory and is eliminated quite quickly as output reverts to its steady trend.

The widespread acceptance of a DS model for aggregate output was based on evidence that the hypothesis of a unit root in real GNP could not be rejected.[1] Although this evidence was extremely robust across various data samples and unit-root testing procedures (see e.g., Charles R. Nelson and Charles I. Plosser, 1982; René M. Stulz and Walter Wasserfallen, 1985; James H. Stock and Mark W. Watson, 1986; Pierre Perron and Peter C. B. Phillips, 1987; Perron, 1988; George W. Evans, 1989), the distinct contrast between traditional TS models of output and recent DS models led many researchers to challenge the unit-root tests. In particular, some questioned the power of these tests, that is, their ability to reject the unit-root null hypothesis when it is indeed

*Board of Governors of the Federal Reserve System, Division of Monetary Affairs, Washington, DC 20551. I thank two anonymous referees for comments as well as Bill Bell, Frank Diebold, Steve Durlauf, Spence Krane, Doug Steigerwald, and David Wilcox. Michelle Phillips and Roberto Sella provided research assistance. The views expressed are my own and are not necessarily shared by the Board of Governors or its staff.

[1]This acceptance was not generally shared by government and business economists. For example, the official series on potential output, which is used in policy and budget deliberations, has remained a smooth nonstochastic trend.

false.[2] However, the indictment of low power against unit-root tests has not been a decisive criticism. The fact that unit-root tests may have low power against certain TS alternatives does not *necessarily* compromise the results from those tests. It would not be at all surprising if unit-root tests had little power against TS alternatives that mimicked a DS model and reverted to trend extremely slowly; all statistical tests have low power against alternatives that are "local" to the null. The failure to reject near-unit-root TS alternatives is of little economic importance, however, because these alternatives are indistinguishable in economic terms from the DS null over the time horizons of practical macroeconomic interest (say, shorter than 10 years). Critics of unit-root tests must instead make the stronger claim that unit-root tests have low power against plausible TS alternatives that display substantially different macroeconomic behavior than a plausible DS null model. Simple power studies are not well suited to answering this question.

This note addresses the argument that unit-root tests have low power only against local alternatives. The goal is to select the most plausible DS and TS representations for output, determine whether these representations have different short-run persistence properties, and then examine whether unit-root tests can distinguish between these models. In the next section, I examine the ordinary least-squares (OLS) estimates of a TS model and a DS model for the postwar sample of real GNP. These models are plausible representations of the data generating process and yet imply very different economic dynamics at the horizons of economic relevance. In Section II, I examine simulated data from these models and show that a unit-root test cannot distinguish between these particular TS and DS representations. Thus, the unit-root test has low power against a plausible TS model that is *not* local in economic terms to a plausible DS model. Section III extends the basic argument to account for small-sample bias in the estimated model coefficients.

I. DS and TS Models of Real GNP

Two obvious candidates for plausible TS and DS representations of the data generating process for real GNP are simply the OLS estimates of these models from the available data sample. The data consist of quarterly observations on U.S. postwar log real GNP per capita (denoted Y_t) from 1948:3 to 1988:4. Assuming second-order dependence,[3] the sample OLS estimate of the TS model of aggregate output with a linear deterministic trend is

$$(1)\quad Y_t = \underset{(0.109)}{-0.321} + \underset{(0.00010)}{0.00030}t + \underset{(0.073)}{1.335}Y_{t-1} - \underset{(0.073)}{0.401}Y_{t-2} + u_t \qquad \hat{\sigma}_u = 0.01013$$

(standard errors of the coefficients appear in parentheses).[4] I will refer to this specific model estimate for the sample as the TS_{OLS} model.

Under the assumption of a unit root, the DS model for this data sample is estimated in first differences as

$$(2)\quad \Delta Y_t = \underset{(0.001)}{0.003} + \underset{(0.074)}{0.369}\Delta Y_{t-1} + \hat{v}_t \qquad \hat{\sigma}_v = 0.01035$$

This particular sample DS model will be denoted as the DS_{OLS} model.

[2]Among many others, Bennett T. McCallum (1986), Francis X. Diebold and Rudebusch (1991), and David N. DeJong et al. (1992) have warned of the low power of unit-root tests. Other challenges have been made by DeJong and Whiteman (1991) from a Bayesian perspective and by Perron (1989) in a framework with structural breaks.

[3]The analysis was repeated assuming that the order of the model was four, six, and eight. Similar persistence and power results were obtained.

[4]The last two quarters of 1948 were used as fixed initial conditions.

The estimated models (1) and (2) both appear to fit real GNP per capita fairly well; the standard deviations of their residuals are quite close, and plots of the residuals suggest no obvious outliers. In addition, Q statistics computed from the fitted residuals provide little evidence against the null hypothesis of no serial correlation at a variety of lags.

However, the estimated TS_{OLS} and DS_{OLS} models have very different implications for the persistence of the dynamic response of output to a random disturbance. To measure this persistence, consider the moving-average representation for the first difference of output implied by a TS or DS model:

$$(3) \quad \Delta Y_t = k + \varepsilon_t + a_1\varepsilon_{t-1} + a_2\varepsilon_{t-2} + \dots$$

where k is some constant and ε_t is the innovation of the model. In this form, the sum of the a_i's measures the model response to a unit innovation.[5] A unit shock in period t affects ΔY_{t+h} by a_h and affects Y_{t+h} by $c_h \equiv 1 + a_1 + \cdots + a_h$. Thus, for various horizons, the cumulative response c_h answers the question: how does a shock today affect the level of real output in the short, medium, and long run? With quarterly data, for example, c_{20} measures the impact of a shock today on Y_t five years hence.

In the limit, the effect of a unit shock today on the level of output infinitely far in the future is given by c_∞. For any TS series, $c_\infty = 0$, because the effect of any shock is eliminated as reversion to the deterministic trend eventually dominates. For a DS series, $c_\infty \neq 0$; that is, each shock has some permanent effect. However, the impulse response of real output at an infinite horizon is of no practical economic significance; indeed, horizons of less than 10 years are usually of greatest interest. At these short horizons, the dynamic responses of TS and DS models may be quite similar or quite different depending on the values taken by the parameters of the models. Thus, the presence of a unit root determines whether c_∞ is positive or zero, but it does not determine all of the model properties of economic interest. It is in this sense that, as noted in the introduction, focusing solely on the existence of unit roots and on the power of unit-root tests against arbitrary TS alternatives is insufficient. What is of economic relevance is the ability of unit-root tests to recognize when data have been generated from TS models that differ substantially from the DS model at short horizons (i.e., the ability to identify economically nonlocal alternatives). Consequently, a comparison of the persistence properties of the estimated TS_{OLS} and DS_{OLS} models at relevant horizons is required.

The estimated model responses are shown in Table 1, with standard errors in parentheses.[6] The impulse response of the DS_{OLS} model implies not only shock persistence but shock magnification. The effect of an innovation is not reversed through time, and it eventually increases the level of real GNP by more than one and a half times the size of the innovation ($c_{20} = 1.59$). In contrast, the TS model exhibits fairly rapid reversion to trend, with 85 percent of a shock dissipated after five years ($c_{20} = 0.15$). Thus, the cumulative impulse responses of these two models, each estimated from the same data sample, imply very different economic dynamics at cyclical frequencies. Because the TS_{OLS} and DS_{OLS} models of aggregate output have such different persistence properties, it would be useful to have a test capable of distinguishing between them. The next section explores the ability of one commonly used unit-root test to accomplish this task.

[5]This measure of persistence is described further in Campbell and Mankiw (1987) and Diebold and Rudebusch (1989).

[6]These standard errors are calculated as follows. Let the cumulative impulse response at horizon h be given by $c_h = F(\rho_1, \dots, \rho_k)$, and let $\mathbf{f}$ denote the vector of partials of F with respect to the parameters; then, the standard error equals $\sqrt{\mathbf{f}'\boldsymbol{\Sigma}\mathbf{f}}$, where $\boldsymbol{\Sigma}$ is the estimated variance-covariance matrix of the autoregressive parameters. The standard errors account for parameter uncertainty but not for unit-root uncertainty.

TABLE 1—CUMULATIVE IMPULSE RESPONSES OF OLS MODELS

Model	Horizon (quarters)								
	1	2	4	8	12	16	20	30	40
DS_{OLS}	1.37	1.51	1.57	1.59	1.59	1.59	1.59	1.59	1.59
	(0.07)	(0.13)	(0.17)	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)
TS_{OLS}	1.33	1.38	1.19	0.73	0.43	0.25	0.15	0.04	0.01
	(0.07)	(0.13)	(0.18)	(0.23)	(0.23)	(0.20)	(0.15)	(0.07)	(0.02)

Note: Standard errors are given in parentheses.

II. Application of a Unit-Root Test

The augmented Dickey-Fuller unit-root test (David A. Dickey and Wayne F. Fuller, 1981) is often used to try to distinguish a TS model from a DS model.[7] For the second-order models under consideration, the augmented Dickey-Fuller regression takes the following form:

$$(4) \quad Y_t = \hat{\mu} + \hat{\gamma} t + \hat{\delta} Y_{t-1} + \hat{\phi}_1 \Delta Y_{t-1} + \hat{\varepsilon}_t .$$

Under the unit-root (or DS model) null hypothesis, $\delta = 1$; thus, the Dickey-Fuller test statistic is simply the t test, $\hat{\tau} \equiv (\hat{\delta} - 1)/\mathrm{SE}(\hat{\delta})$, where $\mathrm{SE}(\hat{\delta})$ is the standard error of the estimated coefficient.

For the postwar real GNP data under consideration, the sample value of the Dickey-Fuller test, which is denoted as $\hat{\tau}_{\mathrm{samp}}$, is equal to -2.98. However, this statistic does not have the usual Student-t distribution, but is skewed toward negative values. At the 10-percent significance level, Dickey and Fuller (1981) calculate the appropriate asymptotic critical value to be -3.12. Thus, the evidence from this sample, in accordance with the findings of previous researchers, suggests that the DS model for real GNP cannot be rejected at even the 10-percent level.

However, the critical values provided by Dickey and Fuller (1981) for their augmented test are only valid asymptotically. In finite samples, the distribution of $\hat{\tau}$ will usually depend on the sample size and nuisance-parameter values (see e.g., Gene Evans and Savin, 1984). These factors can be taken into account by examining simulated data from the DS_{OLS} model and calculating the exact probability of obtaining the sample value of the test statistic from this particular null model. This ensures correct size for the test. More importantly, however, by simulating the TS_{OLS} model, the exact probability of obtaining $\hat{\tau}_{\mathrm{samp}}$ from this particular alternative model can also be obtained. This allows correct assessment of test power against what is arguably one of the most interesting alternatives.

The test-statistic probability distributions conditional on the OLS models are exhibited in Figure 1. The distribution of $\hat{\tau}$ conditional on the DS_{OLS} model is denoted $f_{DS}(\hat{\tau})$, while the distribution of $\hat{\tau}$ conditional on the TS_{OLS} model is denoted $f_{TS}(\hat{\tau})$. Each distribution is formed from 10,000 realizations of the test statistic calculated from 10,000 simulated data samples generated from the particular model.[8] The actual sample value of the test statistic ($\hat{\tau}_{\mathrm{samp}} = -2.98$) is shown as a vertical dotted line.

[7]Similar results to those below were also obtained with the Dickey-Fuller normalized-bias and likelihood-ratio tests, as well as with the generalized Phillips test (Phillips and Perron, 1988).

[8]The samples are generated with normal independently and identically distributed errors with sample size and initial conditions that matched those in equations (1) and (2).

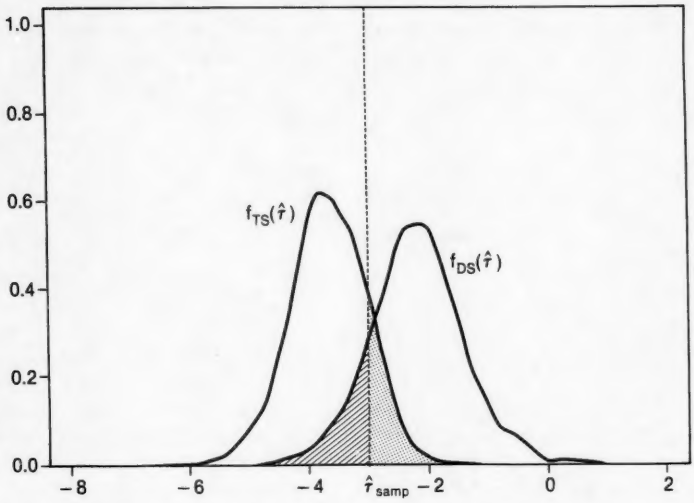


FIGURE 1. EMPIRICAL DENSITIES OF $\hat{\tau}$ FROM DS_{OLS} AND TS_{OLS} MODELS

There are two areas in Figure 1 of special interest. The hatched area under $f_{DS}(\hat{\tau})$ and to the left of $\hat{\tau}_{samp}$ represents the probability of obtaining a value of the t test equal to or smaller than -2.98, conditional on the DS model of equation (2). This p value, is denoted as

$$DS_{OLS}\ p \text{ value} \equiv \text{prob}\left(\hat{\tau} \leq \hat{\tau}_{samp} | DS_{OLS} \text{ model}\right)$$

and represents the marginal significance level for rejection of the null hypothesis for the DS_{OLS} model. This probability equals 0.15; that is, given the sample test statistic, one could not reject the DS model at anything less than the 15-percent level in a classical hypothesis test. This is consistent with the usual inability to reject the DS model for real GNP at conventional significance levels.

The other area of interest is the shaded region under $f_{TS}(\hat{\tau})$ and to the right of $\hat{\tau}_{samp}$. This area represents the probability of obtaining a value of the t test equal to or greater than -2.98, conditional on the TS model of equation (1). This probability is denoted as

$$TS_{OLS}\ p \text{ value} \equiv \text{prob}\left(\hat{\tau} \geq \hat{\tau}_{samp} | TS_{OLS} \text{ model}\right).$$

For real GNP, the TS_{OLS} p value is 0.22, so one would not be able to reject the estimated TS_{OLS} model at even the 20-percent significance level.[9]

In short, the sample statistic for the augmented Dickey-Fuller test does not provide strong evidence against either the estimated DS_{OLS} model or the TS_{OLS} model for real GNP. Earlier papers that are unable to reject a unit root in output provide only one side of the relevant evidence for inference regarding the DS model. The other side, namely, the inability to reject a plausible TS model that exhibits transitory cyclical dynamics of a traditional nature, is at least as convincing.

[9]An equivalent statement of this result is that the $\hat{\tau}$ test at the 15-percent significance level has power against the TS_{OLS} alternative of only 78 percent.

TABLE 2—PROPERTIES OF THE OLS ESTIMATE OF AN AR(1) MODEL

	ρ_1						
Statistic	0.40	0.60	0.80	0.85	0.90	0.95	1.00
$E(\hat{\rho}_1)$	0.38	0.57	0.76	0.81	0.86	0.90	0.94
prob$(\hat{\rho}_1 < \rho_1)$	0.61	0.65	0.72	0.76	0.81	0.89	0.99

Note: Each column is based on 10,000 samples (each with 160 observations) drawn from an AR(1) with an autoregressive coefficient equal to ρ_1.

III. Unbiased DS and TS Models

At first glance, the DS_{OLS} and TS_{OLS} models might appear to be the most plausible candidates for DS and TS representations of the data generating process of real GNP. However, although the OLS estimates of these autoregressive models are consistent and asymptotically normal, they are biased in small samples because the presence of lagged dependent variables violates the assumption of nonstochastic regressors in the classical linear regression model. This bias is easy to illustrate for the OLS estimate of the autoregressive parameter of an AR(1) process,

$$(5) \qquad x_t = \mu + \gamma t + \rho_1 x_{t-1} + \varepsilon_t.$$

Based on 10,000 samples of size 160 generated from equation (5), Table 2 provides the mean value of the OLS estimate, $\hat{\rho}_1$, as well as the proportion of estimates that are less than the true value of ρ_1.[10] For example, if the true ρ_1 is equal to 0.95, the mean OLS estimate is 0.90, and 89 percent of the estimates are less than 0.95. The size of the autoregressive parameter bias that pushes the average OLS estimate below its true value varies with the value of the true parameter, but it is most severe for near-unit-root models (i.e., those with ρ close to 1.0).

A significant bias in the OLS sample estimates is potentially a serious shortcoming of the simulation methodology pursued in the previous two sections. In particular, the estimation bias implies that the sample OLS models used above probably understate the actual amount of persistence in real output. Suppose, for example, that an AR(1) TS representation like equation (5) was fit to real output, and the resulting OLS sample estimate $\hat{\rho}_1$ was equal to 0.90; consequently, the associated estimate of the 10-period cumulative impulse response, $\hat{c}_{10}$, would be 0.34. Assuming that the OLS estimate was equal to its mean, $E(\hat{\rho}_1)$, then the true parameter would be 0.95, and the actual value of c_{10} would be 0.60. Thus, in this case, the OLS sample estimate, on average, understates the amount of persistence.

Arguably, more plausible candidates than the TS_{OLS} and DS_{OLS} models for the data-generating process would correct the parameters for small-sample bias.[11] First-order approximate bias corrections can be calculated for both the TS and DS models quite easily. The bias of the DS_{OLS} model, which is an AR(1) model with an unknown drift parameter, is treated in F. H. C. Marriott and J. A. Pope (1954). They show that, ignoring second-order terms, the expected value of the OLS estimate $\hat{\rho}_1$ is given by

$$(6) \qquad E(\hat{\rho}_1) = \rho_1 - (1+3\rho_1)/T$$

where T is the sample size. Substituting the sample OLS estimate for its expected value

[10] Table 2 is generated with $\mu = \gamma = x_0 = 0$ and disturbances drawn from a standard normal distribution.

[11] In this paper, I consider approximate mean-unbiased estimators of the TS and DS models. In Rudebusch (1992), I examine median-unbiased estimators obtained through repeated simulations, a procedure which would lead to qualitatively similar results if applied to the data set in this paper.

provides a bias-corrected estimator

$$(7) \qquad \tilde{\rho}_1 = (T\hat{\rho}_1 + 1)/(T-3).$$

Note that, to a first-order approximation, this estimator is unbiased:[12]

$$E(\tilde{\rho}_1) = \rho_1.$$

Applying (7) to the DS_{OLS} model for postwar real GNP given in (2), where $\hat{\rho}_1 = 0.369$, I calculate the autoregressive coefficient of the bias-corrected DS model (denoted as the DS_{BC} model) to be $\tilde{\rho}_1 = 0.383$.

Similarly, for the TS_{OLS} model, which is an AR(2) with linear trend, the (first-order) bias in the estimated parameters can be determined using the results in Robert A. Stine and Paul Shaman (1989). Correcting for bias gives a TS_{BC} model with autoregressive coefficients:

$$(8) \qquad \tilde{\rho}_1 = (T\hat{\rho}_1 + 2 + 2\tilde{\rho}_2)/(T-1)$$

for the one-period lag, and

$$(9) \qquad \tilde{\rho}_2 = (T\hat{\rho}_2 + 3)/(T-5)$$

for the two-period lag. For postwar real GNP, the TS_{BC} model has autoregressive coefficients $\tilde{\rho}_1 = 1.351$ and $\tilde{\rho}_2 = -0.395$.

The bias-corrected DS_{BC} and TS_{BC} models both display somewhat greater persistence than the DS_{OLS} and TS_{OLS} models. The bias correction embodied in the DS_{BC} model is in the same direction and of the same magnitude as the one suggested by Table 2. In particular, because the root of the associated lag-operator polynomial is so far from the unit circle, the bias of the DS_{OLS} model is quite modest.[13] For the second-order TS_{OLS} and TS_{BC} models, a useful metric with which to judge their closeness to a nonstationary model is simply the sum of the autoregressive coefficients (see e.g., Phillips, 1991). The sum of the OLS estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ equals 0.933, while the sum of the bias-corrected coefficients $\tilde{\rho}_1$ and $\tilde{\rho}_2$ equals 0.955—a clear, though somewhat small, shift toward nonstationarity.[14]

More specifically, the implications of the bias correction for judging the persistence of real GNP are given in Table 3, which contrasts the impulse responses of the DS_{BC} and TS_{BC} models. The impulse response of the DS_{BC} model implies a shock persistence that is virtually indistinguishable from that of the DS_{OLS} model, which is not surprising given the trivial size of the coefficient bias. Moreover, the TS_{BC} model exhibits reversion to trend only slightly less rapid than that of the TS_{OLS} model. For the TS_{BC} model, almost two-thirds of a shock is dissipated after five years ($c_{20} = 0.37$). Most importantly, it remains true that the DS_{BC} and TS_{BC} candidates for plausible representations of aggregate output have quite different implications about dynamic responses over fairly short horizons.

As a final step, one can ask whether the augmented Dickey-Fuller test can distinguish between these two models. Based on 10,000 samples generated from the DS_{BC} model,[15] I obtain the probability of the sample test statistic as

$$DS_{BC}\ p \text{ value} \equiv \text{prob}\left(\hat{\tau} \leq \hat{\tau}_{\text{samp}} | DS_{BC} \text{ model}\right) = 0.15.$$

In contrast, the probability of the test statistic under the bias-corrected alternative

[12]Of course, reduced bias does not necessarily ensure that $\tilde{\rho}_1$ is a better estimator. However, Guy H. Orcutt and Herbert S. Winokur (1969) explore the properties of this estimator through simulations and find that it often has a smaller mean squared error than the OLS estimator. Theoretical results on this issue are provided in Hong-Ching Zhang (1989).

[13]The bias in the DS_{OLS} model is even smaller than the one in Table 2 because a linear trend is not estimated. With a linear trend, as in Table 2, the bias is given by $E(\hat{\rho}_1) - \rho_1 = -(2 + 4\rho_1)/T$.

[14]The bias correction in the AR(2) TS model may appear to be surprisingly small in light of the large biases shown for the AR(1) model in Table 2. However, as noted by Stine and Shaman (1989), the results for the AR(1) model do not generalize to higher-order models. Indeed, there are cases in which the bias (to a first-order approximation) moves the roots of the lag-operator polynomial *closer* to the unit circle.

[15]The OLS estimates of the trend and intercept are used in generating data.

TABLE 3—CUMULATIVE IMPULSE RESPONSES OF BIAS-CORRECTED MODELS

	Horizon (quarters)								
Model	1	2	4	8	12	16	20	30	40
DS_{BC}	1.38	1.53	1.61	1.62	1.62	1.62	1.62	1.62	1.62
TS_{BC}	1.35	1.43	1.32	0.97	0.70	0.51	0.37	0.16	0.07

model is

TS_{BC} p value

$$\equiv \text{prob}\left(\hat{\tau} \geq \hat{\tau}_{\text{samp}} | \text{TS}_{\text{BC}} \text{ model}\right) = 0.24.$$

These probabilities provide further confirmation of the inability of unit-root tests to identify plausible TS alternative models for real output that display low persistence. In sum, the biases present in the OLS estimates are not substantial enough to change the conclusions from Section I and II.

IV. Conclusion

Based on the usual unit-root tests, little can be said about the relative likelihood of the specific DS and TS models of real GNP given above. In particular, plausible TS alternatives that fit the data cannot be detected because of low test power. The appropriate conclusion from unit-root tests on this data sample is that the existence of a unit root is uncertain.

Furthermore, the unit-root tests employed display low power against plausible TS models that display quite different economic behavior than a plausible DS null model. If the TS models described above correctly portray the persistence of real GNP, then the DS model does not provide a good approximation of the dynamic response of output at even short-run cyclical horizons of, say, five years.

In light of the impulse responses of the TS models, the *appropriate* confidence intervals for estimates of short-term or medium-term persistence are much larger than ones given conditional on the existence of a unit root. This suggests the importance of measuring the confidence intervals for estimates of persistence without conditioning on the TS or DS model. In Diebold and Rudebusch (1989), approximate estimates of such intervals are obtained by using a model of fractional integration that nests the TS and DS models. Stock (1991) also provides a step in this direction by obtaining confidence intervals for the largest autoregressive root.

In sum, the evidence in this paper and in other recent work, notably Lawrence J. Christiano and Martin Eichenbaum (1990), suggests that a new consensus should be formed that stresses the uncertainty about the existence of a unit root in real output and the uncertainty about the amount of persistence of macroeconomic shocks.

REFERENCES

Blanchard, Olivier J., "What is Left of the Multiplier Accelerator?" *American Economic Review*, May 1981 (*Papers and Proceedings*), *71*, 150–4.

Campbell, John Y. and Mankiw, N. Gregory, "Are Output Fluctuations Transitory?" *Quarterly Journal of Economics*, November 1987, *102*, 857–80.

Christiano, Lawrence J. and Eichenbaum, Martin, "Unit Roots in Real GNP: Do We Know, and Do We Care?" *Carnegie-Rochester Conference Series on Public Policy*, Spring 1990, *32*, 7–62.

Clark, Peter K., "The Cyclical Component of U.S. Economic Activity," *Quarterly Journal of Economics*, November 1987, *102*, 797–814.

DeJong, David N., Nankervis, John C., Savin, N. E. and Whiteman, Charles H., "Integration versus Trend Stationarity in Time Series," *Econometrica*, March 1992, *60*, 423–33.

_____ **and Whiteman, Charles H.,** "Reconsidering 'Trends and Random Walks in Macroeconomic Time Series'," *Journal of Monetary Economics*, October 1991, *28*,

221–54.

Dickey, David A. and Fuller, Wayne F., "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root," *Econometrica*, July 1981, *49*, 1057–72.

Diebold, Francis X. and Rudebusch, Glenn D., "Long Memory and Persistence in Aggregate Output," *Journal of Monetary Economics*, September 1989, *24*, 189–209.

_____ **and** _____, "On the Power of Dickey-Fuller Tests Against Fractional Alternatives," *Economics Letters*, 1991, *35* (2), 155–60.

Evans, Gene and Savin, N. E., "Testing for Unit Roots II," *Econometrica* September 1984, *52*, 1241–70.

Evans, George W., "Output and Unemployment Dynamics in the United States: 1950–1985," *Journal of Applied Econometrics*, July-September 1989, *4*, 213–37.

Kydland, Finn and Prescott, Edward C., "A Competitive Theory of Fluctuations and the Feasibility and Desirability of Stabilization Policy," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980, pp. 169–89.

Marriott, F. H. C. and Pope, J. A., "Bias in the Estimation of Autocorrelations," *Biometrika*, December 1954, *41*, 393–403.

McCallum, Bennett T., "On 'Real' and 'Sticky-Price' Theories of the Business Cycle," *Journal of Money, Credit, and Banking*, November 1986, *18*, 397–414.

Nelson, Charles R. and Plosser, Charles I., "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, September 1982, *10*, 139–62.

Orcutt, Guy H. and Winokur, Herbert S., "First Order Autoregression: Inference, Estimation, and Prediction," *Econometrica*, January 1969, *37*, 1–14.

Perron, Pierre, "Trends and Random Walks in Macroeconomic Time Series: Further Evidence from a New Approach," *Journal of Economic Dynamics and Control*, June/September 1988, *12*, 297–332.

_____, "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Econometrica*, November 1989, *57*, 1361–1402.

_____ **and Phillips, Peter C. B.,** "Does GNP Have a Unit Root? A Re-evaluation," *Economics Letters*, 1987, *23* (2), 139–45.

Phillips, Peter C. B., "Bayesian Routes and Unit Roots: De Rebus Prioribus Semper Est Disputandum," *Journal of Applied Econometrics*, October-December 1991, *6*, 435–73.

_____ **and Perron, Pierre,** "Testing for a Unit Root in Time Series Regression," *Biometrika*, 1988, *75* (2), 335–46.

Rudebusch, Glenn D., "Trends and Random Walks in Macroeconomic Time Series: A Re-examination," *International Economic Review*, August 1992, *33*, 661–80.

Stine, Robert A. and Shaman, Paul, "A Fixed Point Characterization for Bias of Autoregressive Estimators," *Annals of Statistics*, 1989, *17* (3), 1275–84.

Stock, James H., "Confidence Intervals for the Largest Autoregressive Root in U.S. Macroeconomic Time Series," *Journal of Monetary Economics*, November 1991, *28*, 435–59.

_____ **and Watson, Mark W.,** "Does GNP Have a Unit Root?" *Economics Letters*, 1986, *22* (2), 147–51.

Stulz, René M. and Wasserfallen, Walter, "Macroeconomic Time-Series, Business Cycles, and Macroeconomic Policies," *Carnegie-Rochester Conference Series on Public Policy*, Spring 1985, *22*, 9–54.

Zhang, Hong-Ching, "Reduction of the Asymptotic Bias of Estimators of Parameters in the Autoregressive and Linear Trend Models," Ph.D. dissertation, Department of Statistics, University of Pennsylvania, 1989.

A Note on Competitive Investment under Uncertainty

By ROBERT S. PINDYCK*

Uncertainty over future output prices or input costs can affect investment by a risk-neutral firm in two opposing ways. First, it can increase the value of the marginal unit of capital, which leads to more investment. This only requires that the stream of future profits generated by the marginal unit be a convex function of the stochastic variable; by Jensen's inequality, the expected present value of that stream is increased. This result was demonstrated by Richard Hartman (1972) and later extended by Andrew Abel (1983) and others. In their models, constant returns to scale and the substitutability of capital with other factors ensure that the marginal profitability of capital is convex in output price and input costs. Even with fixed proportions, however, this convexity can result from the ability of the firm to vary output, so that the marginal unit of capital need not be utilized at times when the output price is low or input costs are high.[1]

If investment is irreversible and can be postponed, a second effect of uncertainty is to create an opportunity cost of investing now, rather than waiting for new information to arrive before committing resources. This increases the full cost of investing in a marginal unit of capital, which reduces investment.[2] Hence the net effect of uncertainty on irreversible investment depends on the size of this opportunity cost relative to the increase in the value of the marginal unit of capital. The sign of the net effect is significant because of its possible policy implications. Pindyck (1988) and Giuseppe Bertola (1989) developed models in which a firm faces a downward-sloping demand curve and found the net effect to be negative.

Ricardo Caballero (1991) has recently studied investment by an individual firm facing a stochastically shifting demand curve. He shows that, as returns to scale become constant and the elasticity of demand faced by the firm rises, this opportunity cost of investing now rather than waiting approaches zero. Thus, in the limit of constant returns and an infinitely elastic demand curve, an increase in uncertainty tends to raise rather than decrease current investment, even if that investment is irreversible.

It is important to stress that Caballero's limiting result treats the firm in isolation and describes the effect of a *mean-preserving* increase in price uncertainty, and not an increase in *industry-wide* demand uncertainty. This note shows that, in a competitive market, the key interactions between irreversibility and uncertainty occur at the industry level and can only be understood by making price and industry output endogenous.[3] Doing so restores the positive opportunity cost associated with irreversible investment.

In addition, Caballero's result is based on a model with convex adjustment costs. An innovative aspect of the model is that these costs can be asymmetric, thereby allowing for partial or complete irreversibility; complete irreversibility corresponds to a cost of

*Massachusetts Institute of Technology, Cambridge, MA 02139. This work was supported by the National Science Foundation under Grant No. SES-8618502, and by MIT's Center for Energy Policy Research. My thanks to Julio Rotemberg and especially Ricardo Caballero for very helpful comments and discussions.

[1]Then the marginal profit of capital at each time t in the future is $\max[0,(p_t - c_t)]$, where c_t is variable cost. Thus, a unit of capital represents a set of all options on future production, which are worth more the greater the variance of p_t or c_t.

[2]For a detailed discussion of this point and a survey of the recent literature on irreversibility and its implications for investment and market evolution, see Pindyck (1991).

[3]Studies that make price endogenous include Steven Lippman and Richard Rumelt (1985), Avinash Dixit (1989, 1991), and John Leahy (1990).

downward adjustment that is infinite. The model is particularly useful in that it allows one to study the sensitivity of investment to the extent of asymmetry in adjustment costs. However, as with other models of this kind, the size of the firm would be unbounded were it not for adjustment costs: it is only adjustment costs that determine firm size. This role of adjustment costs is crucial to the results of Caballero and earlier authors regarding the effect of uncertainty on investment. This note also shows that, while helpful for studying the behavior of a firm in isolation, this adjustment-cost framework (used widely in empirical studies of investment) is inconsistent with a competitive market equilibrium and, hence, with the behavior of a competitive firm.

I. Adjustment Costs, Competition, and Irreversible Investment

A firm that has constant returns to scale everywhere and faces an infinitely elastic demand curve will have a profit function that is linear in the capital stock. Hence, convex costs of some kind are needed to bound the size of the firm; otherwise, the firm would expand indefinitely if the present value of the stream of incremental profits from a unit of capital exceeded the cost of the unit. In most adjustment-cost models of competitive investment, the convexity of adjustment costs limits the size of the firm by making the marginal cost of investment an increasing function of the level of investment. However, because adjustment costs are a function of only the level of investment (and not the stock of capital), investment in each period is independent of investment or the stock of capital in any other period. Hence, there are no linkages between decisions in one period and the next.

This, however, necessarily eliminates irreversibility from the problem. Irreversibility affects today's decisions when it causes those decisions to constrain behavior in the future under some states of nature but not under others. For example, consider the investment decision of a firm facing an uncertain future demand. If the firm invests a large amount this period, it would not want to disinvest next period if demand expands and so would not be constrained by irreversibility. However, the firm would be constrained if demand were to contract next period, because then it would want to disinvest and reduce its capital stock. Because next period's demand is uncertain, irreversibility leads the firm to invest somewhat less this period.

This effect can never arise when the size of the firm is constrained only by adjustment costs. Then, investment next period depends only on the realization of demand that period and on the adjustment-cost function; it is completely independent of investment this period. Hence the firm need only compare the marginal cost of investing to current and expected future marginal profits. If uncertainty increases expected future marginal profits, it will necessarily increase current investment.

Convex adjustment costs may indeed affect the rate at which firms invest (although simple "time to build" and the lumpiness of investment are likely to be more important constraints). It is unrealistic, however, to treat adjustment costs as the sole or main determinant of firm and industry size in equilibrium. In fact, a pure adjustment-cost model with constant returns to scale is inconsistent with a competitive market equilibrium. The reason is that, with free entry, a very large number of very small firms will come into the industry. (Very small firms would enter because they would have very small adjustment costs and hence lower total costs.) In the limit, the industry would be composed of an infinite number of infinitesimally small firms, and so each firm would have *no* adjustment costs.

Even if each firm is constrained to some minimum size, the possibility of entry by new firms or expansion of existing ones will ensure that investment decisions in the face of industry-wide uncertainty are intertemporally linked. As a result, industry-wide uncertainty will affect irreversible investment by a competitive firm with constant returns to scale much as it would a noncompetitive firm or a firm with decreasing returns. The reason is that, in each period, if demand increases existing firms will expand

or new firms will enter until the market clears. From the point of view of an individual firm, this limits the amount that price can rise under good industry-demand outcomes. However, if investment is irreversible, there is no similar mechanism to prevent price from falling under bad demand outcomes. Each firm takes price as given, but it knows that the distribution of future prices is affected by the irreversibility of investment industry-wide. This reduces its own incentive to invest.

Thus, in a competitive equilibrium, uncertainty over market demand affects investment through the feedback of industry-wide capacity expansion and new entry on the distribution of prices.[4] The following example illustrates this with a simple two-period model that allows for this feedback.

II. An Example

As in Caballero's (1991) model, each firm is in place two periods, there is no depreciation or discounting, and the production function is Cobb-Douglas:

$$(1) \qquad q_i = AL_i^{\alpha}K_i^{1-\alpha} \qquad 0<\alpha<1$$

where q_i is the output of firm i. Since I concentrate on competitive firms facing industry-wide shocks, I begin with the *market* demand curve, rather than the demand curve facing an individual firm, and assume that it is isoelastic:

$$(2) \qquad P_t = Q_t^{-1/\varepsilon}Z_t$$

where ε is the elasticity of demand, and Z_t is a stochastic process, with $Z_1 = 1$. For simplicity, let Z_2 equal 0 or 2 with equal probability. I will compare this to the certainty case in which $Z_2 = 1$. Also, I will restrict the discussion to the case of complete irreversibility, with no cost of adjusting K_i upward.

Let there be a large number, N, of equal-size firms, so that each takes price as given, and $Q = Nq_i$. The profit function for each firm is then

$$(3) \qquad \Pi_i = hP^{\eta}K_i$$

where $h = (1-\alpha)A^{1/(1-\alpha)}(\alpha/w)^{\alpha/(1-\alpha)}$ and $\eta = 1/(1-\alpha) > 1$. Also, $q_i = BK_i$, where $B = h/(1-\alpha)$. With no loss of generality, I choose A so that $B = 1$. Note that the value of a marginal unit of capital is hP^{η}, whatever the firm or industry capital stock. This value is convex in P, so its expectation is increased by a mean-preserving spread in P; but as will be seen, this need not mean that uncertainty leads the firm to invest more.

First, consider the certainty case. Here, $P_2 = P_1$, and all investment occurs in period 1. Each firm will want to invest an infinite amount if $2hP_1^{\eta} > k$ and will invest nothing if $2hP_1^{\eta} < k$, where k is the cost of a unit of capital. Thus, in equilibrium, firms invest until price falls to the point that $2hP_1^{\eta} = k$. Hence $P_1 = P_2 = (k/2h)^{1/\eta}$. Industry investment in the first period is $I_1 = K_1 = Q_1 = (2h/k)^{\varepsilon/\eta}$. (Each firm's investment is just $1/N$ of this.)

Now suppose that Z_2 is unknown when firms invest in period 1; it can turn out to be 2 or 0, each with probability 0.5. Although Z_t is exogenous, P_t is determined as part of the market equilibrium. To find this equilibrium, one wants a distribution for P_t that results from Z_t and from firms' investment decisions, with those decisions based on this same distribution.[5]

I will surmise that equilibrium investment in period 1 is small enough so that, in period 2, firms invest some positive amount if $Z_2 = 2$ (whereas they invest nothing if $Z_2 = 0$). After solving for I_1, I will check that this is indeed the case. Then, if $Z_2 = 2$, firms will invest in period 2 to the point that the profit from a unit of capital equals its cost, that is, until $hP_2^{\eta} = k$, or $P_2 = (k/h)^{1/\eta}$. This implies that K_2 will equal

[4]This can also be the case under imperfect competition.

[5]This is easy to do for this simple example. Leahy (1990) solves a more general continuous-time problem.

$(P_2/2)^{-\varepsilon} = 2^{\varepsilon}(h/k)^{\varepsilon/\eta}$. Of course if $Z_2 = 0$, then $P_2 = 0$, $I_2 = 0$, and $K_2 = K_1$.

Given this distribution for price in period 2, risk-neutral firms will invest in period 1 to the point that the expected value of a unit of capital equals its cost:

$$hP_1^{\eta} + E_1(hP_2^{\eta}) = k \tag{4}$$

or

$$hK_1^{-\eta/\varepsilon} + 0.5k = k.$$

Hence $I_1 = K_1 = (2h/k)^{\varepsilon/\eta}$. Finally, I will check that I_2 is indeed positive if $Z_2 = 2$. If $Z_2 = 2$, then

$$I_2 = K_2 - K_1 = (2^{\varepsilon} - 2^{\varepsilon/\eta})(h/k)^{\varepsilon/\eta} > 0$$

since $\eta > 1$.

In contrast to Caballero's (1991) result, I have found that period-1 investment is the same when Z_2 is uncertain as it is when $Z_2 = 1$ with certainty. The reason is that while a mean-preserving spread in the distribution of P_2 increases the value of a unit of capital, a mean-preserving spread in Z_2 reduces the expected value of P_2. The equilibrium response of firms limits price increases under good outcomes of Z_2, but because of the irreversibility of investment, it does not limit price decreases under bad outcomes. In this particular example these two effects just offset each other, so investment is left unchanged.[6]

The Appendix extends this example to n periods and allows Z_t to follow a random walk; it begins at 1 and increases or decreases by 100 percent in each period. The Appendix shows that investment in period 1 is *lower* when Z_t is stochastic, as long as $n \geq 3$. Also, in any period, the difference between investment when future values of Z_t are known and investment when they are stochastic grows with the number of periods remaining. The reason is that the variance of future values of Z_t increases with the time horizon, but industry investment always limits price increases under good outcomes.[7]

III. Concluding Remarks

The simple example presented above shows how industry-wide uncertainty can have a negative effect on irreversible investment, even when firms are perfectly competitive and have constant returns to scale. That effect is mediated by the equilibrium behavior of all firms and the resulting impact on market price. In the two-period example above, that effect just offsets the increase in the value of a unit of capital that results from the convexity of the marginal profit function, so that period-1 investment is left unchanged by a mean-preserving spread in the demand shift variable. When the number of periods exceeds two, period-1 investment is lower when future demand is uncertain.

In the example, I found the equilibrium distribution for price and the levels of investment in each period consistent with that distribution. Alternatively, I could have used the fact, demonstrated by Robert Lucas and Edward Prescott (1971), that the competitive equilibrium is the solution to the social-planning problem. The social planner will use the downward-sloping demand curve to calculate the optimal investment rule and hence will solve an optimal-investment problem that has the same structure as that of a monopolist with constant returns to scale.

The results in this note do not imply that industry-wide uncertainty will have identical effects on investment for a competitive firm and a monopolist. In Caballero and Pindyck (1992), a model with entry and exit is developed that helps clarify the differences between the competitive firm's response to

[6] If $Z_2 = 1$ with certainty, $P_2 = (k/2h)^{1/\eta}$, but if $Z_2 = 0$ or 2, $E_1(P_2) = \frac{1}{2}(k/h)^{1/\eta}$, which is smaller since $\eta > 1$. The expected marginal profit of capital is $k/2$ in both cases.

[7] Like the two-period example, the n-period example in the Appendix ignores depreciation and discounting. Hence if $Z_t = 1$ for all t, the value of a unit of capital grows linearly with n, but it grows less rapidly if Z_t is stochastic and expected future prices are lower. Including depreciation and discounting would reduce the depressive effect of uncertainty on current investment.

industry-wide uncertainty and that of a monopolist and also shows the effects of *firm-level* uncertainty (e.g., over the firm's productivity).

APPENDIX

This appendix extends the two-period example to n periods, where $Z_1 = 1$ and then, in each succeeding period, Z_t increases or decreases by 100 percent, with probability $\frac{1}{2}$ for each. Thus, $Z_2 = 0$ or 2. If $Z_2 = 0$, Z_t remains 0 for all future t, but if $Z_2 = 2$, then $Z_3 = 0$ or 4, and so on. As before, there is no depreciation or discounting. Hence in the certainty case ($Z_t = 1$ always), firms invest in period 1 to the point that

$$\text{(A1)} \qquad nhP_1^{\eta} = nh\left(K_1^{-1/\varepsilon}\right)^{\eta} = k$$

so $I_1 = K_1 = (nh/k)^{\varepsilon/\eta}$, $I_t = 0$ for $t > 1$, and $P_1 = P_2 = \cdots = (k/nh)^{1/\eta}$.

I will again find a solution for the stochastic case by surmising that investment is positive in a good state (i.e., when $Z_{t+1} > Z_t$) and then checking that this is indeed the case. First, in period n, the good state is that in which $Z_n = 2^{n-1}$. In this state, firms invest until $P_n = (k/h)^{1/\eta}$. Thus $K_n = (P_n/Z_n)^{-\varepsilon} = 2^{\varepsilon(n-1)}(h/k)^{\varepsilon/\eta}$, and $I_n = K_n - K_{n-1}$.

In period $n-1$, in the good state $Z_{n-1} = 2^{n-2}$, and firms invest to the point that $hP_{n-1}^{\eta} + E_{n-1}(hP_n^{\eta}) = k$, which implies that

$$\text{(A2)} \qquad h\left(2^{n-2}K_{n-1}^{-1/\varepsilon}\right)^{\eta} + 0.5k = k$$

or

$$K_{n-1} = \left[2^{\eta(n-2)+1}h/k\right]^{\varepsilon/\eta}.$$

Note that since $\eta > 1$, $K_n > K_{n-1}$ in a good state, as was surmised. In period $n-2$, in the good state, firms invest to the point that $hP_{n-2}^{\eta} + E_{n-2}[hP_{n-1}^{\eta} + hP_n^{\eta}] = k$, so that $K_{n-2} = [2^{\eta(n-3)+1}h/k]^{\varepsilon/\eta}$. In general, in a good state,

$$\text{(A3)} \qquad K_{n-m} = \left[2^{\eta(n-m-1)+1}h/k\right]^{\varepsilon/\eta}.$$

Finally, working back to period 1, $I_1 = K_1 = (2h/k)^{\varepsilon/\eta}$. Note that this is smaller than the certainty case when $n > 2$.

REFERENCES

Abel, Andrew B., "Optimal Investment Under Uncertainty," *American Economic Review*, March 1983, *73*, 228–33.

Bertola, Giuseppe, "Irreversible Investment," unpublished manuscript, Princeton University, 1989.

Caballero, Ricardo J., "On the Sign of the Investment–Uncertainty Relationship," *American Economic Review*, March 1991, *81*, 279–88.

_____ **and Pindyck, Robert S.,** "Uncertainty, Investment, and Industry Evolution," National Bureau of Economic Research (Cambridge, MA) Working Paper No. 4160, September 1992.

Dixit, Avinash, "Hysteresis, Import Penetration, and Exchange Rate Pass-Through," *Quarterly Journal of Economics*, May 1989, *104*, 205–28.

_____, "Irreversible Investment with Price Ceilings," *Journal of Political Economy*, June 1991, *99*, 541–57.

Hartman, Richard, "The Effects of Price and Cost Uncertainty on Investment," *Journal of Economic Theory*, October 1972, *5*, 258–66.

Leahy, John V., "Notes on an Industry Model of Entry and Exit," unpublished manuscript, Princeton University, 1990.

Lippman, Steven A. and Rumelt, Richard P., "Industry-Specific Capital and Uncertainty," unpublished manuscript, University of California-Los Angeles, September 1985.

Lucas, Robert E. and Prescott, Edward C., "Investment Under Uncertainty," *Econometrica*, September 1971, *39*, 659–81.

Pindyck, Robert S., "Irreversible Investment, Capacity Choice, and the Value of the Firm," *American Economic Review*, December 1988, *78*, 969–85.

_____, "Irreversibility, Uncertainty, and Investment," *Journal of Economic Literature*, September 1991, *29*, 1110–48.

Bargaining versus Price Competition in Markets with Quality Uncertainty

By HELMUT BESTER*

This paper focuses on the endogenous determination of trading rules. In many markets, as for example in the retail business, prices are simply posted by sellers, and the buyer has little direct influence on how much he has to pay. In other markets prices are the outcome of bilateral negotiations, so that both the seller and the buyer take an active part in setting the price. Examples include not only the bazaar of a less developed nation, but also the markets for used cars, real estate, antiques, and inputs for manufacturing firms. This paper provides a theoretical explanation of which pricing institution is likely to emerge in a market where buyers are imperfectly informed about the quality of goods or services. I compare the performance characteristics of posted-offer pricing with negotiated pricing and find that each arrangement has specific merits. These determine the equilibrium pricing policy as the outcome of competitive interactions between the market participants.

The key insight of my analysis is that the price-determination aspects of market organization can also affect quality. Suppose the buyer has to visit a firm to determine its choice of product quality and that he experiences switching costs when moving from one seller to another. The incentive to provide quality in the posted-price regime is that the consumer may walk away to another store if the quality is too low. Switching costs may undermine this incentive because they create a lock-in effect. A seller who has a locked-in customer may reduce his cost by choosing a lower quality. This argument is often used to advocate self-enforced bans on price advertising for providers of professional services such as doctors and lawyers.[1] While posted-pricing promotes *ex ante* price competition, it may lead to a deterioration in product quality.

In contrast with posted pricing, bargaining determines the price of the good *after* the buyer has arrived at a store and learned its quality. Here the incentive to provide quality is generated by the fact that the seller gets some fraction of the total surplus. He will seek to maximize this surplus by selecting the socially efficient quality. As a result, the negotiated-price market does not exhibit the moral-hazard problem that characterizes the posted-price market. However, the locked-in consumer finds himself in a situation of partial bilateral monopoly with the seller. This allows the seller to exploit his customer, and the bargaining may result in a relatively high price. The different impact of switching costs on price and quality in the posted- and the negotiated-price markets determines the competitiveness of these trading rules.[2]

There is a considerable literature studying the formation of prices in decentralized markets where pairs of agents bargain over the gains from trade.[3] This literature takes

*Center for Economic Research, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. I thank Dilip Mookherjee, Larry Samuelson, Jonathan Thomas, Eric van Damme, and two referees for their comments.

[1] Yuk-Shee Chan and Hayne Leland (1982) and William Rogerson (1988) examine this argument.

[2] For a comparison of posted and negotiated pricing in a laboratory experiment, see James Hong and Charles Plott (1982).

[3] This includes work by Peter Diamond and Eric Maskin (1979), Ariel Rubinstein and Asher Wolinsky (1985), Douglas Gale (1986a, b), and myself (1988). A different context is considered in my 1989 paper, where I replace the price-setting stage of the standard spatial competition model with a bargaining game. Further references and a detailed discussion of bargaining in a market setting are found in the monograph by Martin Osborne and Rubinstein (1990).

the trading rule as exogenously given; the sellers are prohibited from competing with each other by posting prices. The optimal selling strategy of a monopolistic seller is studied by John Riley and Richard Zeckhauser (1983) and by Drew Fudenberg et al. (1987). Their analysis is concerned with the question of whether a fixed posted price yields a higher payoff for the seller than a haggling strategy. Beginning with the work of William Vickrey (1961), the choice of selling mechanisms is an important topic also in the auction literature.[4] It turns out that posting a fixed price is the optimal strategy for a monopolistic seller of a single good when he faces a single potential buyer. Similarly, with many buyers, posted pricing is optimal for the monopolist when he produces the good under constant returns to scale and can freely vary the amount offered for sale (see Milton Harris and Arthur Raviv, 1981a, b). Paul Milgrom (1987) embeds the auction model in a noncooperative-bargaining model in which the owner of the good always has the right to resell the item to other traders. He finds that conducting an auction allows a weak bargainer to benefit from the presence of stronger bargainers, so that typically the sale will be by auction rather than by private offer. The following features distinguish my model from this literature: there are many producers operating under constant returns to scale; these producers can compete by advertising prices; buyers are imperfectly informed about product quality; and buyers face switching costs and cannot negotiate with several sellers at the same time.

To address the problem, I study a simple model that allows one to derive an equilibrium solution both for negotiated and posted pricing. It is presented in Section I. Sections II and III investigate the equilibrium outcome under both trading rules. Based on this analysis, I endogenize the determination of trading rules in Section IV, where I show that for each parameter constellation there is a unique equilibrium pricing mechanism. Concluding remarks are contained in Section V.

[4]See Preston McAfee and John McMillan (1987) and Paul Milgrom (1987) for a survey.

I. The Model

Consider a market with $N > 2$ identical firms. Each firm produces a single good at constant returns to scale. Before the market opens, each firm decides once-and-for-all on the quality $q \in \{q_h, q_\ell\}$ of its output, where $q_h > q_\ell$. The cost of producing one unit of quality q is $c(q)$ with $c(q_h) > c(q_\ell)$. In the model all consumers are identical. They do not interact strategically with each other. This together with the assumption of constant returns to scale allows one to consider each buyer in isolation independently of the total number of consumers. Each consumer purchases at most one unit of the good. His utility from purchasing quality q at the price p is given by $q - p$. Alternatively, he may not purchase the good from any of the N firms and consume some "outside good" instead. The price and the quality of the outside good are exogenously fixed so that the consumer enjoys the net benefit v from buying it.

The buyer does not directly observe the firms' choice of quality. He learns the quality q sold at a particular store only by visiting the store. There is a cost to visiting a store. Switching from one of the N sellers to another or to consuming the outside good takes one time unit. As the buyer discounts future benefits by the discount factor $0 < \delta < 1$, this creates a switching cost. We will view δ as a measure of these costs and investigate its impact on the formation of prices in this market. This is done under the following assumption:

$$(1) \quad q_h - c(q_h) > v > q_\ell - c(q_\ell) > 0.$$

Thus, in the full-information equilibrium with perfect competition all firms would produce quality q_h, and the consumer would buy the high-quality good at the price $p = c(q_h)$. Consuming the outside good would yield a lower utility level. In addition, the

surplus from producing the low-quality good is taken to be too low to compete with the outside good. This implies that under imperfect information the buyer will never visit a store that he suspects to offer quality q_ℓ. Accordingly, I can confine the analysis to situations where the sellers find choosing quality q_h to be optimal.

II. The Negotiated-Price Market

In the negotiated-price market the consumer has the option of purchasing the outside good or visiting one of the N stores to bargain about the price of the good. Upon entering a store, he observes the quality q actually chosen by the seller and so the price negotiations proceed under symmetric information. The "disagreement point" in this bilateral bargaining situation represents the payoffs of the buyer and the seller, respectively, if no sale takes place and the buyer quits; it will be denoted as $(d,0)$. Of course, the buyer's payoff d depends upon his switching cost and the net benefit that he expects from bargaining with another seller or simply from consuming the outside good. Accordingly, in equilibrium, d will be determined endogenously. The seller's profit from not making a sale is zero. He keeps a constant inventory of the good so that after each sale he incurs the cost $c(q)$ of producing one unit. Suppose the generalized Nash bargaining solution is the outcome of the bargaining.[5] Then the price upon which the parties agree is[6]

$$(2)\quad \varphi(q,d) \equiv \operatorname{argmax}_p [q-p-d]^{\alpha} \times [p-c(q)]^{1-\alpha}$$

with $0<\alpha<1$. This solution splits the surplus so that the buyer receives the fraction α. The parameter α may therefore be interpreted as expressing the buyer's "bargaining power"; by varying α from zero to unity, one can obtain any price that is individually rational both for the buyer and the seller. Much of my analysis will focus on the joint impact of the bargaining parameter α and the friction parameter δ on the market outcome.

In the event of breakdown in the negotiations, the buyer can either switch to another bargaining partner or he can purchase the outside good. As the surplus $q_\ell - c(q_\ell)$ is less than v, the buyer will not go to one of the N stores unless he is convinced that he will find quality q_h.[7] In the equilibrium of the negotiated-price market the buyer expects high quality, and he anticipates that the bargaining will result in some price p. Given these expectations and the delay cost of switching, his expected utility from disagreement is

$$(3)\quad d(p) \equiv \max[\delta v, \delta(q_h - p)].$$

In equilibrium the consumer's price–quality expectations have to be consistent with the market outcome.

Definition: $\hat{p}$ is a *negotiated-price equilibrium* if (i) $q_h - \hat{p} \geq v$ and $\hat{p} \geq c(q_h)$; (ii) $\hat{p} = \varphi(q_h, d)$ and $d = d(\hat{p})$; and (iii) $\hat{p} - c(q_h) \geq \varphi(q_\ell, d) - c(q_\ell)$.

The first of these conditions ensures that both the sellers and the buyers are willing to participate in the market. If (i) fails to hold, then none of the N sellers is active, and the consumer purchases the outside good. By (ii), the equilibrium price $\hat{p}$ is determined by the bargaining solution, taking into account that the buyer's threat point

[5]The generalized Nash solution is studied by John Harsanyi and Reinhard Selten (1972). An axiomatization of this solution is given by Ehud Kalai (1977) and Alvin Roth (1979). Kenneth Binmore et al. (1986) derive the generalized Nash bargaining solution as the noncooperative equilibrium of an extensive game in which the parties alternate in making offers and counteroffers.

[6]Of course, an agreement will be reached only if the surplus $q - c(q) - d$ is nonnegative. This condition is always fulfilled in the equilibrium defined below.

[7]As will be shown below, choosing q_h is a dominant strategy for each seller. This precludes the possibility of a mixed equilibrium in which some sellers choose q_h and others choose q_ℓ.

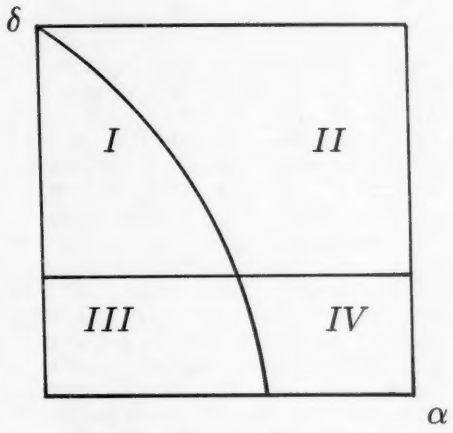


FIGURE 1. NEGOTIATED PRICING IS AN EQUILIBRIUM IN REGIONS II AND IV; POSTED PRICING IS AN EQUILIBRIUM IN REGIONS I AND II

in the price negotiation is $d(\hat{p})$. Finally, (iii) guarantees that each single seller finds it in his own interest to select quality q_h. In equilibrium, the buyer is indifferent among all stores so that the sellers share the market equally. To state the main result of this section, I define the function

$$(4) \qquad \delta_A(\alpha) \equiv \frac{v - \alpha[q_h - c(q_h)]}{(1-\alpha)v}.$$

By assumption (1), $\delta_A(\cdot)$ is decreasing, $\delta_A(0) = 1$, and $\delta_A(\alpha) \geq 0$ for all $\alpha \leq v/[q_h - c(q_h)]$. In Figure 1 the function $\delta = \delta_A(\alpha)$ represents the borderline between the two regions I + III and II + IV.

PROPOSITION 1: *There is a negotiated-price equilibrium if and only if* $\delta \geq \delta_A(\alpha)$. *If* $\delta \geq \delta_A(\alpha)$, *the negotiated-price equilibrium is unique with*

$$\hat{p} = \frac{(1-\alpha)(1-\delta)q_h + \alpha c(q_h)}{1-(1-\alpha)\delta}.$$

PROOF:

The generalized Nash bargaining solution is defined by the necessary and sufficient first-order condition

$$(5) \quad \varphi(q,d) = (1-\alpha)(q-d) + \alpha c(q).$$

Accordingly,

$$\varphi(q,d) - c(q) = (1-\alpha)[q - c(q) - d]$$

so that by assumption (1) equilibrium condition (iii) is satisfied for any $\hat{p}$ satisfying condition (ii). By the first inequality of condition (i), one must have $d(\hat{p}) = \delta(q_h - \hat{p})$. Therefore, solving (ii) for $\hat{p}$ yields the unique solution stated in the proposition. This solution always satisfies the second inequality in (i); the first inequality in (i) is identical to $\alpha[q_h - c_h]/[1-(1-\alpha)\delta] \geq v$. By the definition of δ_A, this is equivalent to $\delta \geq \delta_A(\alpha)$.

The inequality $\delta \geq \delta_A(\alpha)$ is satisfied in regions II and IV of Figure 1. For these parameter constellations, the consumer purchases the high-quality good at the price $\hat{p}$. As $\hat{p}$ exceeds $c(q_h)$, the presence of market frictions enables the sellers to earn positive profits. These are higher, the lower are δ and α. Interestingly, $\hat{p}$ approaches $c(q_h)$ both in the limit when $\delta \to 1$ and in the limit when $\alpha \to 1$. The first of these properties justifies viewing the perfectly competitive outcome as the limiting point of a market with negligible switching costs.

Why does the consumer purchase the outside good for values of δ and α in regions I and III, where a negotiated-price equilibrium fails to exist? The reason is that these parameter values violate equilibrium condition (i). With high switching costs and little bargaining power the buyer cannot get a favorable deal once he has entered a store. Knowing this *ex ante* keeps him from going to a store and induces him to consume the outside good. This kind of market failure is well known from the standard search model (see e.g., Joseph Stiglitz, 1979) in which prices are unilaterally set by sellers, and buyers have to visit stores to observe prices. Indeed, in the limiting situation where $\alpha = 0$ the negotiated-price

market becomes identical to the simplest search model with identical buyers and sellers. When $\alpha = 0$ the seller has all the bargaining power and effectively makes a take-it-or-leave-it offer to his customers. Proposition 1 shows that in this case $\hat{p} = q_h$ (i.e., each seller would charge the consumer's reservation price). This observation is the well-known "monopoly price paradox" of Diamond (1971): even with arbitrarily small search costs and a large number of firms, in market equilibrium the price is the monopoly price. If the buyer has to pay a cost for entering the market or can opt for some outside good, this paradox predicts total market failure.[8] In regions II and IV this unsatisfactory outcome is avoided because α is large enough to guarantee the buyer a sufficient share of the gains from trade.

Importantly, the proof of Proposition 1 reveals that the negotiated-price market involves no problem of moral hazard; that is, the incentive constraint (iii) is never binding. This perhaps surprising observation has a simple intuition: the bargaining outcome guarantees the seller a fraction $(1 - \alpha)$ of the bargaining surplus. As a result, he is always better off by producing q_h because this quality yields a higher surplus than does q_ℓ. Choosing quality q_h is a dominant strategy for the seller in the negotiated-price market.[9] This fact distinguishes this market from the posted-price market, which I turn to in the next section, where prices are set *before* the consumer becomes aware of qualities.[10]

III. The Posted-Price Market

In the posted-price market, the sellers act as Bertrand competitors by posting prices. Advertising price information enables the seller to guarantee his customers a price before they visit his store. In contrast we assume that communication of quality information is infeasible. This assumption is relevant in markets where quality is either too costly to communicate or is not verifiable to third parties. If a court finds it hard to determine whether a seller actually provides "high" quality, then buyers must distrust quality advertisements because they cannot be enforced.

The buyer observes the sellers' price quotations and compares their attractiveness with the outside-option utility v. After entering a store and learning its quality he can either make a purchase at the posted price or switch to another seller. By assumption (1), he will not go to a store if he anticipates finding quality q_ℓ.[11] In the posted-price equilibrium the buyer expects q_h in each store, and so all sellers post the same price p^*. As all stores appear identical to the buyer, he visits one of them at random. To confirm his expectations, competition must induce the suppliers to offer quality q_h.[12] As q is not directly observable, each single seller has an incentive to select q_h at the posted price p^* only if the buyer would quit after observing quality q_ℓ. Given his expectations about qualities at other stores, the buyer will certainly do so if $q_\ell - p^* < d(p^*)$. In fact, we will assume that the buyer refuses to purchase low quality unless he is not actually worse off by departing.[13] This means

[8] In the literature, the typical way to avoid the Diamond result has been to assume that some consumers are well informed about prices (see Steven Salop and Stiglitz, 1977; Hal Varian, 1980; Dale Stahl, 1989).

[9] This conclusion is robust against affine transformations of the seller's and buyer's utilities. The reason is that the generalized Nash bargaining solution satisfies the axiom of independence of utility rescalings.

[10] The relationship between *ex ante* and *ex post* pricing in a model without qualitative uncertainty is explored by Gale (1988).

[11] Of course, the buyer presumes that no seller offers the good at a price below cost.

[12] I assume that each seller selects q_h when he is indifferent between q_h and q_ℓ.

[13] This tie-breaking rule is necessary to avoid the open-set problem that would occur if there were no lowest price that signals high quality. Indeed, equilibrium condition (iii) in the next (posted-price equilibrium) definition is impossible to satisfy if condition (ii) is replaced by $p^* > \pi(p^*)$.

that each single seller will choose quality q_h if and only if the equilibrium price p^* satisfies the restriction $p^* \geq \pi(p^*)$, where

$$\pi(p^*) \equiv q_\ell - d(p^*). \tag{6}$$

While $p^* \geq \pi(p^*)$ ensures the provision of high quality, a key factor in the determination of p^* is that prices are a signal of quality. Should some seller deviate from p^* by posting $p < p^*$, then the buyers will use the observed price to draw inferences about this seller's quality. If they interpret p as a signal of quality q_ℓ, they will not be attracted even though $p < p^*$. The opposite happens if p is regarded as a signal of quality q_h. As in other signaling games, such an indeterminacy of out-of-equilibrium beliefs may lead to a multiplicity of equilibrium prices p^*. To avoid this problem, I will restrict the buyers' beliefs to satisfy the "intuitive criterion" proposed by In-Koo Cho and David Kreps (1987). Suppose some seller wants to undercut his competitors by some price p slightly below p^* and, at the same time, wishes to convince the consumer that he offers high quality. Then one may reasonably assume that this seller succeeds if he would not gain by posting p and selecting quality q_ℓ, even if his offer attracts the entire market. This prerequisite is fulfilled if low quality deters the customer from paying p for the good, that is, if $p \geq \pi(p^*)$. Summing up, in the posted-price equilibrium p^* only prices $p \geq \pi(p^*)$ are considered as a signal of high quality.

Definition: p^* is a *posted-price equilibrium* if (i) $q_h - p^* \geq v$ and $p^* \geq c(q_h)$; (ii) $p^* \geq \pi(p^*)$; and (iii) there is no $p \geq \pi(p^*)$ such that $p < p^*$ and $p - c(q_h) > [p^* - c(q_h)]/N$.

The first of these conditions is the same as in the definition of the negotiated-price equilibrium. Requirement (ii) represents the sellers' incentive-compatibility constraint to provide high quality. Condition (iii) precludes any of the sellers gaining by unilaterally posting some price p below p^* that signals high quality. Here we assume that if all sellers post the same price p^*, each store has the same chance of attracting consumers so that its market share equals $1/N$. The equilibrium outcome depends on the level of switching costs. Let

$$\delta_B \equiv 1 - (q_h - q_\ell)/v. \tag{7}$$

Then, $\delta_B < 1$ because $q_h > q_\ell$.

PROPOSITION 2: *There is a posted-price equilibrium if and only if* $\delta \geq \delta_B$. *If* $\delta \geq \delta_B$, *the posted-price equilibrium is unique with* $p^* = \max[c(q_h), (q_\ell - \delta q_h)/(1-\delta)]$.

PROOF:

By the first inequality in equilibrium condition (i), one has $\pi(p^*) = q_\ell - \delta(q_h - p^*)$. As $p^* \geq c(q_h)$ and $N \geq 2$, condition (iii) is satisfied if and only if p^* minimizes p subject to $p \geq c(q_h)$ and $p \geq \pi(p^*)$. For $\delta \geq [q_\ell - c(q_h)]/[q_h - c(q_h)]$, only the first constraint is binding and so one has $p^* = c(q_h)$. Otherwise, only the second constraint is binding so that $p^* = \pi(p^*)$, that is, $p^* = (q_\ell - \delta q_h)/(1-\delta)$. If $p^* = c(q_h)$, the first inequality in (i) is always satisfied. For $p^* > c(q_h)$ this inequality becomes $q_h - p^* = (q_h - q_\ell)/(1-\delta) \geq v$ which, by definition of δ_B, is identical to $\delta \geq \delta_B$.

The posted-price equilibrium may fail to exist for low values of δ if $\delta_B > 0$, that is, if

$$q_h - q_\ell < v. \tag{8}$$

This condition implies that the consumer prefers buying the outside good to purchasing quality q_h at a price $p > q_\ell$. It limits the use of prices as signals of quality in the posted-price market. Even though the buyer may reasonably be convinced that prices above q_ℓ indicate high quality, because he would always quit a low-quality store with $p > q_\ell$, he cannot be attracted by such a price offer. More generally, the lock-in effect becomes more serious when the difference between q_h and q_ℓ is decreased. This is so because the consumer's utility gain from quitting a low-quality seller and

going to a high-quality seller becomes smaller.[14]

Regions I and II of Figure 1 describe the area where a posted-price equilibrium exists. For values of δ below δ_B the lock-in effect becomes too strong and leads to a deterioration of product quality. As a consequence, the institution of posted-offer pricing precludes the sellers from being active in the market and results in consumption of the outside good in regions III and IV.

For δ close enough to unity, the posted price p^* equals $c(q_h)$. For lower values of δ, however, one observes that price exceeds marginal costs. The increase in the price above the cost of producing high quality gives the firms an incentive not to offer low quality. This is a familiar feature of search models in which prices signal product quality (see Asher Wolinsky, 1983; Rogerson, 1988). A similar mechanism appears also in experience good markets with repeat purchases, in which quality is learned after the good is bought. In the Benjamin Klein and Keith B. Leffler (1981) model, consumers pay a quality premium which prevents the seller from cutting quality. Michael Riordan (1986) assumes that firms make price and quality decisions which cannot be altered over the relevant period. A high price then signals a high quality since consumers would refuse a *second* purchase of low quality at a high price. In my model, the consumer learns about quality before buying the good and so he *never* purchases low quality at a high price.

The negotiated-price market is clearly more efficient than the posted-price regime in region IV of Figure 1. Here the N firms remain inactive in the posted-price market, whereas negotiated pricing results in production of the high-quality good with positive payoffs both for the sellers and the buyers. In contrast, the posted-price market appears to be superior in region I, where the buyer refrains from entering negotiations because his bargaining power is too low. The key insight from Propositions 1 and 2 is that the two categories of trading institutions involve a trade-off. Price bargaining avoids the moral-hazard problem in the firms' selection of qualities; yet, as the price is determined *ex post* after the buyer has chosen the seller, it may not guarantee the buyer a sufficient fraction of the surplus to make bargaining attractive *ex ante*. *Ex ante* pricing, as in the posted-price market, does not suffer from this drawback; but, when the price is fixed *ex ante*, the lock-in effect may have a negative impact on the seller's incentive to produce high quality. The relative importance of these considerations depends on the parameters δ and α. The following section will demonstrate that the trade-off between the two pricing institutions can explain which will emerge as an equilibrium trading rule in a given environment.

[14]The same argument shows that the equilibrium quality must unravel to q_ℓ when quality is a continuous choice variable within some interval $[q_\ell, q_h]$. To avoid total market failure in this situation, one has to assume that some consumers are well informed about quality.

IV. The Stability of Competition

This section is devoted to analyzing which pricing mechanism is stable against competition. A particular trading rule can survive only if no trader can gain by deviating and using another trading rule. Applying this idea to the negotiated-price market means that no seller should be able to profit from posting a price *ex ante* in a situation where all the other sellers rely on *ex post* pricing. That is, it must be impossible to attract profitably the demand of all consumers by undercutting the negotiated-price equilibrium $\hat{p}$ and posting $p < \hat{p}$. The reason why such an attempt may fail is that prices below $\hat{p}$ may be viewed as an indication of low quality. Using the same restrictions on beliefs as in Section III, I will assume that the posted offer p convinces the consumer of quality q_h only if $p \geq \pi(\hat{p})$, where $\pi(\cdot)$ is defined by (6).

Definition: The negotiated-price equilibrium $\hat{p}$ is *stable* against price competition if there is no $p < \hat{p}$ such that (i) $p \geq \pi(\hat{p})$ and (ii) $p - c(q_h) > [\hat{p} - c(q_h)]/N$.

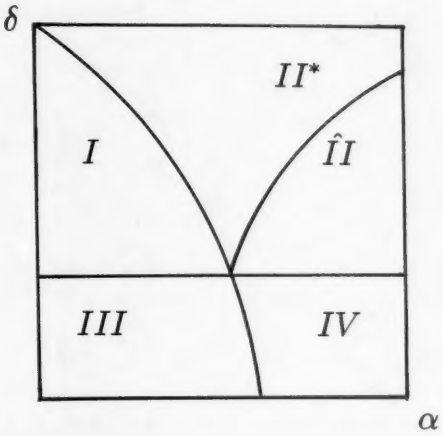


FIGURE 2. NEGOTIATED PRICING IS STABLE IN REGIONS $\widehat{\text{II}}$ AND IV; POSTED PRICING IS STABLE IN REGIONS I AND II*

In other words, the institution of negotiated pricing cannot be eroded by price-posting if any offer below $\hat{p}$ is either viewed as a low-quality signal or fails to increase the seller's profit even when he serves the whole market. To determine the range of parameter values for which this is true, I define the function

(9)
$$\delta_C(\alpha) \equiv \frac{\alpha(q_h - c(q_h)) - (q_h - q_\ell)}{\alpha(q_h - c(q_h)) - (1-\alpha)(q_h - q_\ell)}.$$

Note that for

$$\alpha^* \equiv [q_h - q_\ell]/[q_h - q_\ell + q_h - c(q_h) - v]$$

one has

$$\delta_A(\alpha^*) = \delta_B = \delta_C(\alpha^*).$$

Moreover, $\delta_C'(\alpha) > 0$ for all $\alpha \in (\alpha^*, 1)$. In Figure 2 the function $\delta = \delta_C(\alpha)$ is depicted for $\alpha \in (\alpha^*, 1)$; it divides the former region II of Figure 1 into the regions II* and $\widehat{\text{II}}$.

PROPOSITION 3: *The negotiated-price equilibrium $\hat{p}$ is stable against price competition if and only if $\alpha \geq \alpha^*$ and $\delta \leq \delta_C(\alpha)$.*

PROOF:

As $\hat{p} > c(q_h)$ and $N \geq 2$, there is always a $p < \hat{p}$ satisfying (ii) in the definition of stability. As $q_h - \hat{p} \geq v$, condition (i) is identical to $p \geq q_\ell - \delta(q_h - \hat{p})$. Therefore, $\hat{p}$ is stable if and only if there is no $p < \hat{p}$ satisfying (i). This means one must have $\hat{p} \leq q_\ell - \delta(q_h - \hat{p})$. Using $\hat{p}$ from Proposition 1, this condition is equivalent to

$$(10) \quad \delta[\alpha(q_h - c(q_h)) - (1-\alpha)(q_h - q_\ell)] \leq \alpha(q_h - c(q_h)) - (q_h - q_\ell).$$

As $\delta < 1$, this inequality cannot hold if the left-hand side is negative. Accordingly, by the definition of $\delta_C(\alpha)$, (10) holds if and only if $\alpha > [q_h - q_\ell]/[q_h - c(q_h) + q_h - q_\ell] \equiv \bar{\alpha}$ and $\delta \leq \delta_C(\alpha)$. Note that $\delta_C(\alpha)$ is strictly increasing for $\alpha > \bar{\alpha}$ and that $\delta_A(\alpha^*) = \delta_C(\alpha^*)$. Accordingly, by Proposition 1 there is no negotiated-price equilibrium $\hat{p}$ for $\alpha \in (\bar{\alpha}, \alpha^*)$ and $\delta \leq \delta_C(\alpha)$. This means that condition (10) applies if and only if $\alpha \geq \alpha^*$ and $\delta \leq \delta_C(\alpha)$.

Proposition 3 states that negotiated pricing cannot be sustained as a Nash equilibrium in the firms' choice of pricing policies for parameter constellations in regions I and II* of Figure 2. Negotiated pricing constitutes a stable equilibrium only in regions $\widehat{\text{II}}$ and IV. At first sight it may appear paradoxical that for $\delta < \delta_C(1)$ the sellers will rely on price bargaining when the buyer's bargaining power is rather high. This is so, however, because competition forces sellers to adopt a trading rule that is advantageous for the buyer.

Interestingly, notice that a stable negotiated price market necessitates a certain amount of market frictions. As $\delta_C(1) < 1$, bargaining is not a stable pricing institution when δ is close to unity. Negotiated pricing is unlikely to survive in a highly competitive, almost frictionless environment. The empirical implication is that bargaining tends to be replaced by posted pricing when im-

provements in the communication and transportation technology reduce the buyer's search cost. Switching costs in combination with imperfect quality information provide a role for negotiated pricing in regions $\widehat{\text{II}}$ and IV. This may explain why there is relatively little haggling in markets where the consumer is well informed about product quality as, for example, in the market for mass-produced brand goods. Also, bargaining is observed less frequently in the market for new automobiles than in the used-car market, where quality uncertainty is more important. Other examples for negotiated pricing are the markets for antiques and real estate, where the buyer has to inspect quality before making a purchase.

To complete the analysis, I now investigate the stability of posted pricing. I assume that posting p^* legally commits the seller only in the sense that he cannot ask his customers to pay more than p^*. However, this does not constrain the parties not to revise jointly the terms of the transaction. If in the course of bargaining they both reach an agreement, then this replaces the posted price. Accordingly, the buyer accepts the seller's posted offer p^* only if he does not see a chance to pay less after bargaining.

Definition: The posted-price equilibrium p^* is *stable* against bargaining if, given $d = d(p^*)$, it is the case that $\varphi(q_h, d) \geq p^*$.

Given that the buyer cannot induce a price reduction by bargaining, he has to pay p^* after switching to another store. Therefore, his threat point in a stable posted-price equilibrium is $d(p^*)$, as defined by (3).

PROPOSITION 4: *The posted-price equilibrium p^* is stable against bargaining if and only if either $\alpha \leq \alpha^*$ or $\alpha > \alpha^*$ and $\delta \geq \delta_C(\alpha)$.*

PROOF:

Using (5), p^* is stable if and only if $p^* \leq (1-\alpha)[q_h - d(p^*)] + \alpha c(q_h)$. As $d(p^*) = \delta(q_h - p^*)$, this is equivalent to

$$(11)\quad p^* \leq \left[(1-\alpha)(1-\delta)q_h + \alpha c(q_h)\right] / \left[1-(1-\alpha)\delta\right].$$

By Proposition 2 this condition always holds if $\delta \geq [q_\ell - c(q_h)]/[q_h - c(q_h)]$ because then $p^* = c(q_h)$. For $\delta \in (\delta_B, [q_\ell - c(q_h)]/[q_h - c(q_h)])$, one has $p^* = (q_\ell - \delta q_h)/(1-\delta)$, so that (11) is identical to

$$(12)\quad \delta\left[\alpha(q_h - c(q_h)) - (1-\alpha)(q_h - q_\ell)\right] \geq \alpha(q_h - c(q_h)) - (q_h - q_\ell).$$

As $0 < \delta < 1$, (12) always holds if the left-hand side is negative, that is, if

$$\alpha \leq [q_h - q_\ell] / [q_h - c(q_h) + q_h - q_\ell] \equiv \bar{\alpha}.$$

For $\alpha > \bar{\alpha}$, (12) is equivalent to $\delta \geq \delta_C(\alpha)$. By Proposition 2, p^* exists if and only if $\delta \geq \delta_B$. As $\delta_B \geq \delta_C(\alpha)$ for $\alpha \in (\bar{\alpha}, \alpha^*)$, any p^* is stable if $\alpha \leq \alpha^*$. For $\alpha \in [\alpha^*, 1)$, one has $\delta_B \leq \delta_C(\alpha) < [q_\ell - c(q_h)]/[q_h - c(q_h)]$, so that over this range (11) holds if and only if $\delta \geq \delta_C(\alpha)$.

Propositions 3 and 4 demonstrate that, whenever the N sellers are active in the market, a unique stable pricing institution emerges. As the stability criterion eliminates posted pricing in region $\widehat{\text{II}}$ of Figure 2, our model predicts that Bertrand competition will prevail in regions I and II* and negotiated pricing will prevail in regions $\widehat{\text{II}}$ and IV. Using Propositions 1 and 2, it is easily established that $p^* < \hat{p}$ in region II* whereas $\hat{p} < p^*$ in region $\widehat{\text{II}}$. The endogenous determination of trading rules thus maximizes the consumer's equilibrium utility. In region II* the sellers are trapped in a prisoner's dilemma type of situation. They all end up with lower profits because the negotiated price $\hat{p}$ makes undercutting profitable. In contrast, in region $\widehat{\text{II}}$ the signaling effect associated with posted pricing results in a price level p^* that makes bargaining more efficient for coping with the moral-hazard problem.

V. Conclusion

I have explored how different pricing mechanisms affect the determination of quality and price in a market with quality uncertainty and switching costs. In sum-

mary, posted pricing involves moral hazard, whereas negotiated pricing is not very competitive. I have shown that in my model this trade-off between the two trading institutions uniquely determines the equilibrium pricing policy. The equilibrium trading mechanism has the interesting property that it ensures the consumer the highest possible utility level.

My model stresses the role of quality uncertainty for the determination of pricing rules. Of course, this leaves out a number of other considerations that may be important. For instance, I have assumed that bargaining proceeds under symmetric information so that negotiations are costless. Asymmetric-information bargaining models can generate costs in the form of delay in agreement.[15] This may favor posted pricing. In general, however, it is not clear a priori which is the most efficient pricing institution when information and incentive problems are involved.

REFERENCES

Admati, Anat and Perry, Motty, "Strategic Delay in Bargaining," *Review of Economic Studies*, July 1987, *54*, 345–64.

Bester, Helmut, "Bargaining, Search Costs and Equilibrium Price Distributions," *Review of Economic Studies*, April 1988, *55*, 201–14.

_____, "Noncooperative Bargaining and Spatial Competition," *Econometrica*, January 1989, *57*, 97–113.

Binmore, Kenneth G., Rubinstein, Ariel and Wolinsky, Asher, "The Nash Bargaining Solution in Economic Modelling," *Rand Journal of Economics*, Summer 1986, *17*, 176–88.

Chan, Yuk-Shee and Leland, Hayne, "Prices and Qualities in Markets with Costly Information," *Review of Economic Studies*, October 1982, *49*, 499–516.

Cho, In-Koo and Kreps, David M., "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, May 1987, *102*, 179–221.

Diamond, Peter A., "A Model of Price Adjustment," *Journal of Economic Theory*, June 1971, *3*, 156–68.

_____ **and Maskin, Eric,** "An Equilibrium Analysis of Search and Breach of Contract, I: Steady States," *Bell Journal of Economics*, Spring 1979, *10*, 282–316.

Fudenberg, Drew, Levine, David and Tirole, Jean, "Incomplete Information Bargaining with Outside Opportunities," *Quarterly Journal of Economics*, February 1987, *102*, 37–50.

Gale, Douglas, "Bargaining and Competition, Part I: Characterization," *Econometrica*, July 1986, *54*, 785–806.

_____, "Bargaining and Competition, Part II: Existence," *Econometrica*, July 1986, *54*, 807–18.

_____, "Price Setting and Competition in a Simple Duopoly Model," *Quarterly Journal of Economics*, November 1988, *103*, 729–39.

Gul, Faruk and Sonnenschein, Hugo, "On Delay in Bargaining with One-Sided Uncertainty," *Econometrica*, May 1988, *56*, 601–11.

Harris, Milton and Raviv, Arthur, (1981a) "Allocation Mechanisms and the Design of Auctions," *Econometrica*, November 1981, *49*, 1477–99.

_____ **and** _____, (1981b) "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *American Economic Review*, June 1981, *71*, 347–65.

Harsanyi, John and Selten, Reinhard, "A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information," *Management Science*, January 1972, *18*, 80–106.

Hong, James T. and Plott, Charles R., "Rate Filing Policies for Inland Water Transportation: An Experimental Approach," *Bell Journal of Economics*, Spring 1982, *13*, 1–19.

Kalai, Ehud, "Nonsymmetric Nash Solutions and Replications of Two-Person Bargaining," *International Journal of Game The-*

[15]The simplest model of this type assumes one-sided uncertainty about the buyer's valuation of the good and considers an extensive game in which the seller makes a sequence of offers to the buyer as, for example, in Joel Sobel and Ichiro Takahashi (1983). For the role of delay in bargaining, see also Anat Admati and Motty Perry (1987) and Faruk Gul and Hugo Sonnenschein (1988).

ory, 1977, *6* (3), 129–33.

Klein, Benjamin and Leffler, Keith B., "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, August 1981, *89*, 615–41.

McAfee, Preston and McMillan, John, "Auctions and Bidding," *Journal of Economic Literature*, June 1987, *25*, 699–738.

Milgrom, Paul R., "Auction Theory," in Truman Bewley, ed., *Advances in Economic Theory*, Cambridge: Cambridge University Press, 1987, pp. 1–32.

Osborne, Martin J. and Rubinstein, Ariel, *Bargaining and Markets*, San Diego: Academic Press, 1990.

Riley, John and Zeckhauser, Richard, "Optimal Selling Strategies: When To Haggle, When To Hold Firm," *Quarterly Journal of Economics*, May 1983, *98*, 267–89.

Riordan, Michael H., "Monopolistic Competition with Experience Goods," *Quarterly Journal of Economics*, May 1986, *101*, 265–79.

Rogerson, William P., "Price Advertising and the Deterioration of Product Quality," *Review of Economic Studies*, April 1988, *55*, 215–29.

Roth, Alvin, *Axiomatic Models of Bargaining*, Berlin: Springer-Verlag, 1979.

Rubinstein, Ariel and Wolinsky, Asher, "Equilibrium in a Market with Sequential Bargaining," *Econometrica*, September 1985, *53*, 1133–50.

Salop, Steven and Stiglitz, Joseph, "Bargains and Rip-offs: A Model of Monopolistically Competitive Price Dispersion," *Review of Economic Studies*, October 1977, *44*, 393–406.

Sobel, Joel and Takahashi, Ichiro, "A Multi-Stage Model of Bargaining," *Review of Economic Studies*, July 1983, *50*, 411–26.

Stahl, Dale O., "Oligopolistic Pricing with Sequential Consumer Search," *American Economic Review*, September 1989, *79*, 700–12.

Stiglitz, Joseph E., "Equilibrium in Product Markets with Imperfect Information," *American Economic Review*, May 1979 (*Papers and Proceedings*), *69*, 339–45.

Varian, Hal R., "A Model of Sales," *American Economic Review*, September 1980, *70*, 651–9.

Vickrey, William, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance*, March 1961, *16*, 8–37.

Wolinsky, Asher, "Prices as Signals of Product Quality," *Review of Economic Studies*, October 1983, *50*, 647–58.

Rigging the Lobbying Process: An Application of the All-Pay Auction

By MICHAEL R. BAYE, DAN KOVENOCK, AND CASPER G. DE VRIES*

Why do politicians frequently "announce" that they have narrowed down a set of potential recipients of a "prize" to a slate of finalists?[1] In general, does the slate of finalists comprise the "best" candidates, and does the best candidate always win?[2] This paper provides answers to these questions. Our model of the political process is one of rent-seeking, which takes the (perhaps overly jaded) view that persons with power award political prizes on the basis of self-interest.

In a world where a politician can explicitly auction off a prize to the high bidder, the standard auction literature can be used to analyze political behavior. The justice system, however, precludes politicians from explicitly selling the prize to the highest bidder; thus politicians cannot let it become public knowledge that they are in the business of selling political favors.

An interesting institution has emerged in political markets to overcome this constraint: lobbying. Lobbyists make *implicit* payments to the politician, through campaign contributions or "wining-and-dining." If these up-front payments were rebated to those failing to receive the prize, it would be clear that the politician was selling favors. It is natural, therefore, for a political institution to arise such that lobbyists "ante up" before the prize is awarded, and these up-front payments are not refunded to those failing to win the prize. This view of lobbying has a structure isomorphic to the all-pay auction, which differs from standard auctions in one principal respect: all bids are forfeited by the bidders.

Before we describe our model of the lobbying process, it is useful to provide an overview of the existing literature and to contrast it with the present analysis. The case in which more than two lobbyists value the prize identically was first analyzed by Hervé Moulin (1986), who characterizes the symmetric equilibrium to the all-pay auction. Similar analysis is provided by Arye Hillman (1988), who argues that the equilibrium is unique. It turns out, however, that the symmetric equilibrium is not unique; in fact there is a *continuum* of equilibria (in Baye et al. [1990], we provide a full characterization of the equilibria.)

The case in which some lobbyists value the prize more than others has been analyzed by, among others, Hillman and John Riley (1989), who argue that equilibrium involves only the top two lobbyists. In this *Review*, Tore Ellingsen (1991) has considered the interesting case in which one lobbyist values the prize more than $n-1$ competitors with common valuations (see his proposition 1), and he demonstrates the existence of n equilibria. Baye et al. (1990) have shown, however, that there actually exists a continuum of equilibria in this case. Moreover, the expected revenue earned by the politician differs across this continuum of equilibria; there is *not* revenue equivalence across the equilibria.

The present analysis provides a simple closed-form expression for expected revenues that is valid for *all* equilibria. Our

*Baye: Department of Economics, Pennsylvania State University, University Park, PA 16802; Kovenock: Department of Economics, Purdue University, West Lafayette, IN 47907; de Vries: Department of Economics, Erasmus Universiteit Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. We are grateful to Donald Deere, Tom Gresik, Tim Gronberg, Bill Nielson, Guoqiang Tian, Steve Wiggins, and three anonymous referees for suggestions that led to improved content and exposition.

[1]The International Olympic Committee, for instance, selected six cities as "finalists" for the 1996 Summer Olympics: Belgrade, Manchester, Toronto, Melbourne, Athens, and Atlanta.

[2]Atlanta won the bid for the 1996 Olympics.

technique does not require an explicit calculation of the Nash-equilibrium mixed strategies. Consequently, our results provide a framework with which one may reexamine the implications of the equilibria missed for previous results, without explicitly calculating the (uncountable infinity) of equilibria.

The objective of the present paper is to take into account the continuum of equilibria and to determine the amount of rents the politician can expect to earn given such a political institution. To answer this question, we model the political process as a two-stage game of complete information. In stage 1 the politician takes the political institution of lobbying as given but is free to constrain the process by "narrowing down" the slate of candidates to a set of "finalists." In stage 2, the finalists compete in an all-pay auction: the lobbyist giving the greatest bribe wins the prize, while the others receive nothing for their payments. We solve this decision problem by backwards induction, solving first for the Nash-equilibrium payments that accrue to the politician in the second-stage lobbying game given an arbitrary set of lobbyists. We then solve for the optimal first-stage decision of the politician, which involves the selection of the set of "finalists" that maximizes expected political rents. We will show that, under plausible circumstances, the politician has a perverse incentive to preclude lobbyists most valuing the prize from participating in the second-stage lobbying game. Intuitively, this precommitment may take the form of announcing prior to any lobbying that "five states have been selected as finalists for the site of a new military base." We will refer to this precommitment as the *exclusion principle*. The exclusion principle has obvious implications for efficiency; states deriving the greatest economic benefit from a military base (and hence having the highest valuation of the prize) may be excluded a priori from the announced set of finalists.

I. The Model

Consider a politician who must determine which of $n > 2$ lobbyists will receive a prize. The value of the prize to lobbyist i is $v_i > 0$, where the v_i's are common knowledge and ordered such that $v_1 \geq v_2 \geq \cdots \geq v_n$.

The politician does not care which lobbyists wins the prize but does care about how much money he has available in his campaign chest. Accordingly, he decides to award the prize to the lobbyist who gives him the greatest up-front, nonrefundable *implicit* bribe. The objective of the politician is to select a set of lobbyists (the set of finalists) that maximizes his expected rents, $W = E\sum_{i=1}^{n} b_i$, where b_i is the bribe paid by player i.

Given a set of "finalists," lobbying is an all-pay auction: the payoff to lobbyist i if he offers a bribe of b_i is $\pi_i = v_i - b_i$ if b_i is the highest of all n bribes. However, if some other lobbyist offers a higher bribe, lobbyist i's payoff is $\pi_i = -b_i$. We assume that when multiple lobbyists submit the highest bribe, the prize is awarded to one of them at random. Thus, the payoff of lobbyist i is given by

$$(1)\quad \pi_i(b) = \begin{cases} v_i - b_i & \text{if } b_i > b_j \ \forall\, i \neq j \\ \dfrac{v_i}{M} - b_i & \text{if } i \text{ ties } M-1 \text{ others for high bid} \\ -b_i & \text{if } b_i < b_j \text{ for some } j \neq i. \end{cases}$$

This payoff structure is standard in the lobbying literature (cf. Hillman, 1988). It can also be viewed as the limiting case of an alternative payoff structure suggested by Gordon Tullock (1980) that is also used in this literature (see Baye et al., 1989). In Section II we characterize the expected payments by lobbyists in the second-stage lobbying game. These results are used in Section III to determine the politician's rent-maximizing selection of the set of "finalists."

II. The Lobbying Game

We first sketch a proof of the nonexistence of a pure-strategy Nash equilibrium for the all-pay auction. Consider the two-player case and suppose (b_1, b_2) did comprise a pure-strategy Nash equilibrium

(without loss of generality, suppose $b_1 \geq b_2$). If $b_1 < v_2$, player 2 could deviate to earn a higher payoff by increasing b_2 slightly above b_1 to win the prize. If $b_1 \geq v_2$, player 2's best reply to b_1 is zero; but with $b_2 = 0$, it pays player 1 to deviate from b_1 by lowering the bid to (small) $\varepsilon > 0$, contradicting the hypothesis that $b_1 \geq v_2$.

It is known, though, that there does exist an equilibrium in mixed strategies, in which lobbyists randomize their bribes (cf. Partha Dasgupta and Eric Maskin, 1986; Moulin, 1986; Hillman and Riley, 1989; Baye et al., 1990). Moreover, with more than two players, there generally exists a continuum of possible equilibria (Baye et al., 1990). As our focus centers around the politician's rent-maximizing selection of finalists, we need only characterize the expected total bribes that accrue in a given Nash equilibrium of the lobbying game. The innovation is that the techniques employed below do not rely on the algebraic form of the mixed strategies used by the lobbyists in equilibrium and, thus, are valid even in the presence of a continuum of Nash equilibrium mixed strategies.

The following theorem is the key ingredient that enables us to determine the set of finalists that maximizes the politician's rents. The novelty of the result is that it is valid for *each* equilibrium in the continuum of possible Nash equilibria and thus can be used for purposes beyond the present paper. For example, the formula allows one to strengthen the results of Ellingsen (1991), which are based on a finite subset of equilibria.

THEOREM 1: *Let $v_1 \geq v_2 \geq \cdots \geq v_n$ denote the valuations of lobbyists $\{1,2,\ldots,n\}$ in the stage-2 lobbying game. Let $E_1 b_1$ denote the expected bid of a lobbyist with the highest valuation. Then in any Nash equilibrium,*

$$(2) \quad W = \frac{v_2}{v_1} v_2 + \left[1 - \frac{v_2}{v_1}\right] E_1 b_1 \leq v_2.$$

PROOF:

Let $F_i(b_i)$ denote the cumulative distribution function of lobbyist i in an arbitrary (mixed-strategy) Nash equilibrium, and let S_i denote the support of the distribution. Lobbyist i must earn constant (expected) profits almost everywhere (a.e.) in S_i. For lobbyist 1 this constant must equal $v_1 - v_2$, and for lobbyists $2,3,\ldots,n$, this constant is zero (see Baye et al., 1990).[3] Hence, the following conditions must hold:

$$(3) \quad \pi_1(b_1) = \prod_{i \neq 1}^{n} F_i(b_1)[v_1 - b_1] + \left[1 - \prod_{i \neq 1}^{n} F_i(b_1)\right][-b_1] = v_1 - v_2 \quad \text{a.e. on } S_1$$

and

$$(4) \quad \pi_i(b_i) = \prod_{j \neq i}^{n} F_j(b_i)[v_i - b_i] + \left[1 - \prod_{j \neq i}^{n} F_j(b_i)\right][-b_i] = 0 \quad \text{a.e. on } S_i,\ i \neq 1.$$

Let $p_i(b_i) \equiv \prod_{j \neq i}^{n} F_j(b_i)$ denote the probability that lobbyist i wins the prize, conditional on his bid and the strategies employed by the other $n-1$ lobbyists in a Nash equilibrium.[4] Then, since equations (3) and (4) hold almost everywhere in their respective supports, taking the expectations of these equations and manipulating reveals that

$$(5) \quad P_1 v_1 - E_1 b_1 = v_1 - v_2$$

and

$$(6) \quad P_i v_i - E_i b_i = 0 \qquad \forall\, i \neq 1$$

where E_j denotes the expectation with respect to lobbyist j's (equilibrium) mixed

[3] Note that when lobbyist 2's valuation equals that of lobbyist 1, $v_1 - v_2 = 0$.

[4] We can rule out mass points for any agent at a bid $b > 0$ (see Baye et al., 1990).

strategy and $P_j \equiv E_j p_j(b_j)$. Summing over equations (5) and (6), we then obtain

$$(7) \quad W \equiv \sum_{j=1}^{n} E_j b_j = (P_1 - 1)v_1 + \sum_{i \neq 1} P_i v_i + v_2.$$

Applying the fundamental theorem of integral calculus to $\sum_{j=1}^{n} P_j$, it follows that $\sum_{j=1}^{n} P_j = 1$. Furthermore, if $v_2 > v_i$, $i > 2$, then $P_i = 0$ (see Baye et al., 1989, 1990). Hence,

$$(8) \quad W = (P_1 - 1)v_1 + (2 - P_1)v_2.$$

Rearranging (5), we find

$$P_1 = \frac{v_1 - v_2 + E_1 b_1}{v_1}$$

which, inserted into (8), yields our results.

Two implications of Theorem 1 are worth noting. First, if two or more players most value the prize at some common level, v, the expected rents accruing to the politician equal v; there is full rent dissipation. Secondly, if $v_1 > v_2$, then the expected rents accruing to the politician are strictly *less* than v_2, since $E_1 b_1 < v_2$ in any Nash equilibrium. In other words, regardless of whether there is a unique equilibrium[5] or a continuum of equilibria,[6] in every equilibrium there is *under*dissipation of rents. In the following section, this result will be used to establish when it pays a politician to preclude some lobbyists from competing in the lobbying game. First, however, we state the following lemma from Hillman and Riley (1989).

LEMMA 1: *Suppose that the valuations of the lobbyists in the stage-2 lobbying game are such that* $v_1 \geq v_2 > v_3 \geq v_4 \cdots \geq v_n$. *Then in the unique Nash equilibrium,* $Eb_1 = v_2/2$.

[5]The equilibrium is unique when $v_2 > v_3$.
[6]There is a continuum of equilibria when $v_2 = v_3$.

Theorem 1 and Lemma 1 together imply that, when two players value the prize strictly more than all other players, the expected rents accruing to the politician in the stage-2 lobbying game are

$$(9) \quad W(v_1, v_2) = \left[1 + \frac{v_2}{v_1}\right]\frac{v_2}{2}.$$

Note that, when $v_1 > v_2 > v_3$, expected rents are increasing in v_2 but decreasing in v_1. Intuitively, as player 1's valuation increases, the playing field becomes more unequal. Hence, player 2 reduces his expected payment to the politician, and total expected rents decline.

It is important to note that the formula in equation (9) is based on specific configurations of valuations and does not hold in general (it does not hold when $v_1 > v_2 = v_3$). The reason is that, when $v_1 > v_2 = v_3$, Eb_1 in equation (2) varies depending upon which of the continuum of equilibria is played, and thus, the politician's expected rents depend upon which equilibrium the lobbyists play. This point has not been addressed in the existing literature, and it plays a crucial role in our analysis.

III. Selecting the Finalists

Since there exists a continuum of expected political rents for some configurations of valuations, our next task is to characterize properties of the maximum expected political rents that can be extracted from the lobbyists.

PROPOSITION 1: *If* $\{\hat{1}, \ldots, \hat{m}\}$ *is a rent-maximizing set of finalists (with valuations* $\hat{v}_1 \geq \cdots \geq \hat{v}_m$*), then expected rents are*

$$(10) \quad W(\hat{v}_1, \hat{v}_2) = \left(1 + \frac{\hat{v}_2}{\hat{v}_1}\right)\frac{\hat{v}_2}{2}.$$

PROOF:

We must show that if $\{\hat{1}, \ldots, \hat{m}\}$ is a set of finalists that maximizes expected rents (and the corresponding valuations are $\hat{v}_1 \geq \cdots \geq$

$\hat{v}_m$), then expected rents are $W(\hat{v}_1, \hat{v}_2)$. This is clearly true if $m = 2$; hence, suppose $m > 2$. If $\hat{v}_1 = \hat{v}_2 \equiv \hat{v}$, equation (2) reveals that $W = \hat{v} = W(\hat{v}_1, \hat{v}_2)$. If $\hat{v}_1 > \hat{v}_2 > \hat{v}_3$, equation (9) shows again that $W = W(\hat{v}_1, \hat{v}_2)$. Finally, if $\hat{v}_1 > \hat{v}_2 = \hat{v}_3 \equiv \hat{v}$, expected rents increase by excluding player $\hat{1}$, since by Theorem 1 $W(\hat{v}_1, \hat{v}_2) < \hat{v}_2 = W(\hat{v}_2, \hat{v}_3)$. However, this contradicts the hypothesis that the set $\{\hat{1}, \ldots, \hat{m}\}$ maximizes expected rents. Hence, we conclude that any rent-maximizing set of finalists generates expected rents of $W(\hat{v}_1, \hat{v}_2)$.

Thus, while equation (9) does not hold for all possible configurations of values, it does hold when the set of finalists is selected so as to maximize expected rents [equation (10)]. This result allows us to determine the set of finalists that maximizes the politician's expected rents. Specifically, since equation (10) is decreasing in the highest valuation and increasing in the second-highest valuation, it never pays to exclude a player with a valuation that lies between the valuations of any two lobbyists who are in the set of finalists. Thus, the expected rent-maximizing set of finalists is determined by considering all pairwise combinations of adjacent lobbyists until lobbyists k and $k+1$ are found such that

$$W(v_k, v_{k+1}) = \max_i W(v_i, v_{i+1}).$$

To realize these rents, the politician must exclude players with valuations greater than v_k from the set of finalists. Formally, we have shown the following:

PROPOSITION 2: *Suppose* $v_1 \geq v_2 \geq v_3 \geq \cdots \geq v_n$. *Then the politician maximizes expected rents by constructing a set of finalists that excludes lobbyists with valuations strictly greater than* v_k, *where k is such that*

$$\left(1 + \frac{v_{k+1}}{v_k}\right)\frac{v_{k+1}}{2} \geq \left(1 + \frac{v_{i+1}}{v_i}\right)\frac{v_{i+1}}{2} \quad \forall\, i.$$

In order to highlight the implications of our results, consider the following two corollaries.

COROLLARY 1: *Suppose* $v_1 = v_2 \geq v_3 \cdots \geq v_n$. *Then the politician does not gain by constructing an agenda that excludes some lobbyists from the lobbying game.*

COROLLARY 2: *Suppose* $v_1 > v_2 = v_3 \geq \cdots \geq v_n$. *Then the politician maximizes expected rents by excluding the lobbyist with the highest valuation from the set of finalists.*

It may also be optimal for a politician to exclude more than one lobbyist from the stage-2 game. For instance, suppose $v_1 > v_2 > v_3 = v_4 \geq \cdots \geq v_n$. Then the politician maximizes expected rents by constructing an agenda that excludes lobbyists 1 and 2 from the set of finalists whenever

$$\left(1 + \frac{v_2}{v_1}\right)\frac{v_2}{2} < v_3.$$

These results demonstrate the *exclusion principle*: a politician may benefit from precluding the lobbyists valuing the prize the most from participating in the lobbying process.

We conclude with a numerical example to aid in elucidating our findings. Suppose $v_1 = 50$, $v_2 = 40$, and $v_3 = 38$. The theorem and lemma imply that the politician earns $W = 36$ if he does not constrain the lobbying process or limits lobbying to only players 1 and 2. However, if the politician announces that players 2 and 3 are the finalists, then the expected payments to the politician are $W(40, 38) = 37.05$. Thus it pays the politician to exclude lobbyist 1, who values most the prize, from participating in the lobbying game.

IV. Conclusions

This paper has examined an interesting principle arising in all-pay auctions: the exclusion principle. This principle states that a politician wishing to maximize political rents may find it in his best interest to exclude certain lobbyists from participating in the lobbying process—particularly lobbyists valuing most the political prize. In addition to pointing out the exclusion principle, our

Theorem 1 characterizes expected revenue for the entire continuum of equilibria that can arise in $n > 2$-player all-pay auctions with arbitrary valuations of the prize. This is in contrast to the results of Ellingsen (1991), Hillman (1988), Hillman and Riley (1989), and Hillman and Dov Samet (1987), among others, which are valid only for a subset of possible equilibria.

REFERENCES

Baye, Michael R., Kovenock, Dan and de Vries, Casper G., "The Economics of All-Pay, Winner-Take-All Contests," Texas A&M University Working Paper No. 89-21, 1989.

_____, _____ **and** _____, "The All-Pay Auction with Complete Information," CentER Working Paper No. 9051, Tilburg University, 1990.

Dasgupta, Partha and Maskin, Eric, "The Existence of Equilibrium in Discontinuous Economic Games, I: Theory," *Review of Economic Studies*, January 1986, *53*, 1–26.

Ellingsen, Tore, "Strategic Buyers and the Social Cost of Monopoly," *American Economic Review*, June 1991, *81*, 648–57.

Hillman, Arye L., *The Political Economy of Protectionism*, New York: Harwood, 1988.

_____ **and Riley, John,** "Politically Contestable Rents and Transfers," *Economics and Politics*, Spring 1989, *1*, 17–39.

_____ **and Samet, Dov,** "Dissipation of Contestable Rents by Small Numbers of Contenders," *Public Choice*, 1987, *54* (1), 63–82.

Moulin, Hervé, *Game Theory for the Social Sciences*, 2nd Ed., New York: New York University Press, 1986.

Tullock, Gordon, "Efficient Rent Seeking," in J. M. Buchanan, R. D. Tollison, and G. Tullock, eds., *Toward a Theory of the Rent Seeking Society*, College Station: Texas A&M University Press, 1980, pp. 97–112.

Monopolistic Competition and Optimum Product Diversity: Comment

By XIAOKAI YANG AND BEN J. HEIJDRA*

The model of Avinash Dixit and Joseph E. Stiglitz (1977) (DS hereafter) has received more and more attention from economists in the past decade and can now be regarded as the central model of monopolistic competition. Applications of the model can be found in such diverse fields as international trade, growth theory, and macroeconomics.[1]

The purpose of this short note is twofold. First, we suggest an alternative solution method to that proposed by DS and used in most applications of their model. Our method makes full use of the symmetry assumption employed by DS but extends their method, so that the resultant pricing rule takes into account the price-index effect. Our method retains the intuitive appeal and simplicity of the original DS model. Second, we show that our solution method extends the range of applicability of the DS model significantly; more general production structures can now be incorporated in the model.

The note is organized as follows. In Section I we set up the DS model and develop our solution method. We also discuss the impact that the alternative solution method has on the qualitative results obtained in the DS study and the early application by Paul R. Krugman (1980). Section II broadens the scope of the discussion somewhat by discussing existence problems associated with the DS approach under increasing returns and isoelastic cost functions. It is shown that if our solution method is used then the existence problems disappear, and the DS model yields well-defined answers. Finally, in Section III some concluding remarks are given.

I. The Preference-for-Diversity Model

Consider the following version of the DS model.[2] Consumers maximize utility subject to a budget constraint. The utility function is separable in the numeraire commodity x_0 and the other commodities $(x_i,\ i=1,\ldots,n)$. The prices of the commodities are denoted by $p_0\ (=1)$ and p_i, respectively. Income is equal to the endowment income (I) plus distributed profits. Since profit is zero in equilibrium, income equals I. The maximization problem is:

$$(1)\qquad \max_{x_0,x_i} u = U\left(x_0, \left[\sum_{i=1}^{n} x_i^{\rho}\right]^{1/\rho}\right)$$

$$\text{subject to}\qquad I = \sum_{i=0}^{n} p_i x_i.$$

The price and quantity indexes q and y are

*Yang: Reader, Department of Economics, Monash University, Clayton, Victoria, Australia 3168; Heijdra: Research Fellow, OCFEB, Erasmus University, Postbus 1738, 3000 DR Rotterdam, The Netherlands. We thank the referees as well as the following people for their helpful comments: Yew-Kwang Ng, Peter Dixon, Avinash Dixit, and Gene Grossman.

[1] In the international-trade literature the seminal papers are Paul R. Krugman (1980), Wilfred J. Ethier (1982), and Gene M. Grossman and Elhanan Helpman (1989, 1990). A recent growth model employing the DS insights is Paul M. Romer (1990). The main macroeconomic applications are Blanchard and Kiyotaki (1987) and Julio J. Rotemberg (1987).

[2] We follow the notation introduced by DS with the exceptions that we do not set endowment income I equal to unity and that we use σ to denote the substitution elasticity between the differentiated commodities. The example discussed in the text refers to the constant-elasticity case of section I of DS. This is also the case discussed by Jean Tirole (1988 pp. 298–300).

defined as follows:

$$(2) \qquad q \equiv \left[\sum_{i=1}^{n} p_i^{1-\sigma} \right]^{1/(1-\sigma)}$$

$$y \equiv \left[\sum_{i=1}^{n} x_i^{\rho} \right]^{1/\rho}.$$

Here $\sigma \equiv 1/(1-\rho)$ is the elasticity of substitution between the differentiated products. Since $0 < \rho < 1$, it follows that $\sigma > 1$. Part a of Proposition 1 summarizes the solutions of this maximization program.[3] Demands are uniquely determined by the share function, $s(q)$, the form of which depends on U.[4]

PROPOSITION 1: (*a*) *The solutions for* x_0 *and* x_i *are*

$$(3) \qquad x_i = I \frac{s(q)}{q} \left(\frac{q}{p_i} \right)^{\sigma}$$

$$x_0 = [1 - s(q)] I.$$

(*b*) *The own elasticity of demand faced by firm i is denoted by* ε_i $[\equiv (\partial x_i / \partial p_i)(p_i / x_i)]$, *and the cross elasticity of demand for firm i is denoted by* ε_{ij} $[\equiv (\partial x_i / \partial p_j)(p_j / x_i)]$. *In the symmetric equilibrium with* $p_i = p_j \ \forall i, j$, *these elasticities are as follows:*

$$(4) \qquad \varepsilon_i = -\sigma + \frac{\sigma - 1}{n} + \frac{\theta(q)}{n}$$

$$\varepsilon_{ij} = \varepsilon_i + \sigma.$$

Here $\theta(q)$ $[\equiv (\partial s / \partial q)(q / s)]$ *is the elasticity of the share function.*

Equation (3) is the DS solution. The elasticity of demand (ε_i) as given in (4), however, differs from that calculated by DS, who find $\varepsilon_i = -\sigma$. The DS approach makes two assumptions to approximate a solution:

ASSUMPTION 1: *Each producer i ignores the cross-price elasticity of demand for a variety of goods (i.e.,* $\varepsilon_{ij} = 0$*).*

ASSUMPTION 2: *The influence of an individual price change on the general price index is ignored (i.e.,* $\partial q / \partial p_i = 0$*).*

Assumption 1 is ad hoc and is accepted here since no alternative is available that is both tractable and theoretically superior.[5] However, as we argue below, in any application of this model one must ascertain *ex post* that ε_{ij} is indeed close to zero *in equilibrium*.

Assumption 2 is not innocuous. DS motivate this assumption on the grounds that the number of goods is extremely large. This claim may be criticized on the following grounds. First, the number of goods in the DS model is *endogenously* determined and is a function of technology and endowment parameters. Hence, we need some qualifications on these parameters to ensure that the number of goods is indeed large in equilibrium. We cannot simply assume that this condition is satisfied. Second, as we demonstrate below, the effects on the model's solution of ignoring the price index in the calculation of the demand elasticity may not be trivial even if the number of goods is large, depending on the relative values of some key parameters.

Assumptions 1 and 2 are in a sense both ad hoc, but there is nevertheless virtue in relaxing Assumption 2 and not Assumption 1. In the setting adopted here, and from a

[3] Proofs of propositions have been omitted for reasons of space. They are available from the authors upon request.

[4] For example, if U is Cobb-Douglas with parameter $1-\gamma$ for the numeraire good, the $s(q) = \gamma$.

[5] One admittedly ad hoc way to introduce perceived price reaction is to postulate $\partial p_j / \partial p_i = \xi$ for all $i \neq j$, where $0 \leq \xi \leq 1$. Then, the price elasticity in symmetric equilibrium is equal to

$$\varepsilon_i = -\sigma + [1 + \xi(n-1)](\sigma - 1 + \theta)/n.$$

This provides an additional rationale for a term involving n in the elasticity formula. See Tirole (1988 pp. 244–5) for a critique of this approach.

purely technical point of view, Assumptions 1 and 2 are related: $\varepsilon_{ij} = 0 \Leftrightarrow \varepsilon_i = -\sigma$. If we adopt Assumption 1, then Assumption 2 is implied and vice versa. This equivalence does not validate forcing this result by assuming that n is large; a large n must be the outcome, not an assumption of the analysis. From an economic point of view, Assumptions 1 and 2 are furthermore quite distinct: Assumption 1 relates to behavior external to the firm, while Assumption 2 is internal to the firm. If we abandon Assumption 1 we leave the world of monopolistic competition. This is not the case with Assumption 2.

A further issue cuts deeper at the heart of the DS approach. A key assumption of that model is that the elasticity of substitution σ is constant; no matter how many varieties are available, the substitution elasticity is constant. A more reasonable assumption may be that σ rises with the number of available varieties: $\sigma = \sigma(n)$ with $\sigma' > 0$. Clearly, in such a case ε_{ij} will only tend to zero as $n \to \infty$ if $\sigma(n)$ rises at a slower rate than n. Otherwise, a large n may still be associated with a nontrivial cross elasticity.

The elasticity of demand as given in (4) incorporates the price-index effect (in symmetric equilibrium) without significantly complicating the solution of the model. We now investigate whether our solution method affects any of the qualitative conclusions reached by DS. Several equilibria can be distinguished; (i) the *market equilibrium* (ME) is the case of decentralized Chamberlinian monopolistic competition, (ii) the *constrained social optimum* (CSO) is defined as the optimum that would be obtained by a benevolent social planner in the absence of lump-sum subsidies, and (iii) the *unconstrained social optimum* (USO) is the optimum when lump-sum subsidies are allowed.

ME occurs when the following two conditions are satisfied by all active firms i:

$$p_i = c\left(\frac{\varepsilon_i}{1+\varepsilon_i}\right) \tag{5}$$

$$(p_i - c)x_i = a.$$

The first part of (5) implies equality of marginal revenue and marginal cost (MR = MC) for each active producer, where c is marginal cost. The second part of (5) ensures that the marginal firm breaks even, where a is fixed cost. In symmetric equilibrium all firms charge the same price and produce the same amounts (i.e., $p_i = p_j = p$ and $x_i = x_j = x$).

Explicit solutions for p, x, and n cannot be obtained for the general case defined here. This is irrespective of the solution method used. DS characterize the solution qualitatively under the assumption that ε_i is constant [i.e., by ignoring the terms involving n from the expression for ε_i in (4)]. They show that the equilibrium is unique, provided the following elasticity condition is satisfied: $\theta(q) > 1 - \sigma$ (DS, 1977 p. 300). In principle, this condition could be generalized to take into account the expanded expression for ε_i in (5). We have chosen, instead, to base our comparison of the solution methods on a special case, where U in (1) is Cobb-Douglas. In that case, $s(q) = \gamma$, and $\theta(q) = 0$. Our rationale for concentrating on this special case is twofold. First, the major applications of the DS model have been based on this special case (see e.g., Krugman, 1980; Blanchard and Kiyotaki, 1987). Second, simple and intuitive explicit results are possible using this assumption.[6]

Part a of Proposition 2 summarizes the ME solutions for the equilibrium price, output, and the number of products. For purposes of comparison, parts b and c of Proposition 2 restate the results of DS.

PROPOSITION 2: *Assume that U is Cobb-Douglas. Then $s(q) = \gamma$, $\theta(q) = 0$.*

[6]Assuming a constant-elasticity-of-substitution (CES) formulation for U still does not render $\theta(q)$ constant and hence does not lead to explicit solutions. See equation 6 in DS (p. 298) for the relationship among the substitution elasticity, the share function, and its elasticity.

(a) The market equilibrium (ME) solutions for n, $x_i = x$, and $p_i = p$ are:

$$(6) \qquad p_e = \frac{c}{\rho}\left(\frac{\gamma I}{\gamma I - a}\right)$$

$$x_e = \frac{\rho(\gamma I - a)}{c\left[\frac{\gamma I}{a}(1-\rho)+\rho\right]}$$

$$n_e = \frac{\gamma I}{a}(1-\rho)+\rho.$$

(b) The constrained social optimum (CSO) solutions for n, $x_i = x$, and $p_i = p$ are:

$$(7) \qquad p_c = \frac{c}{\rho}$$

$$x_c = \frac{a\rho}{c(1-\rho)}$$

$$n_c = \frac{\gamma I}{a}(1-\rho).$$

(c) The unconstrained social optimum (USO) solutions for n, $x_i = x$, and $p_i = p$ are:

$$(8) \qquad p_u = c$$

$$x_u = \frac{a\rho}{c(1-\rho)}$$

$$n_u = \frac{\gamma I(1-\rho)}{a[\rho+\gamma(1-\rho)]}.$$

(d) The degree of diversification in ME, CSO, and USO may be compared as follows:

$$(9) \qquad n_e > n_c \quad n_u > n_c$$

$$\rho \in [0, \tilde{\rho}) \Rightarrow n_u > n_e$$

$$\rho = \tilde{\rho} \Rightarrow n_u = n_e$$

$$\rho \in (\tilde{\rho}, 1] \Rightarrow n_u < n_e$$

where

$$\tilde{\rho} = \tilde{\rho}(a, \gamma; I) \quad \partial\tilde{\rho}/\partial a < 0$$

and

$$\partial\tilde{\rho}/\partial\gamma < 0$$

$$\tilde{\rho}(a, 1; I) = 0.$$

(e) In ME the own and cross price elasticities are equal to the following:

$$(10) \qquad \varepsilon_i = \frac{-1}{1-\rho+\frac{a\rho}{\gamma I}}$$

$$\varepsilon_{ij} = \frac{\rho}{(1-\rho)\left[\frac{\gamma I}{a}(1-\rho)+\rho\right]}.$$

DS themselves find $x'_e = a\rho/c(1-\rho)$, $p'_e = c/\rho$, and $n'_e = \gamma I(1-\rho)/a$. Comparing these values to the ones in (6) reveals that DS underestimate the number of firms (although by less than 1, since $0 < \rho < 1$), underestimate the price, and overestimate the quantity. The difference in price disappears, however, if one assumes (as do DS) that n'_e represents a large number. Of course, in view of the definition for n'_e this will only be large if fixed costs (a) are low relative to the share of endowment income spent on the differentiated commodities (γI) and if the goods are not close to perfect substitutes ($\rho \ll 1$). If the value of ρ is close to 1 then it all depends on the relative magnitude of $(1-\rho)/a$ even if γI is large.

A more subtle consequence of our correction to the DS solution method appears when we compare the ME solutions to those corresponding with the constrained and unconstrained social optima as defined by DS (1977 pp. 300–2). Part b of Proposition 2 summarizes the CSO solutions for the equilibrium price, output, and the number of products. Comparing the CSO solutions to the ME solutions reveals that the market equilibrium does indeed lead to "a little too

much" diversification (i.e., $n_e > n_c$). By using our solution method one finds that the conclusion reached by DS (1977 p. 301) regarding the equivalence of ME and CSO no longer holds.

Part c of Proposition 2 summarizes the USO solutions for the equilibrium price, output, and the number of products. As in DS, price equals marginal cost, and the consumers pay the lump-sum subsidies to the producers to render profits equal to zero.

Part d of Proposition 2 summarizes the results regarding the degree of diversification under the different scenarios. As in DS (1977 p. 302) USO yields a greater degree of diversification than the CSO (i.e., $n_u > n_c$). However, *unlike* DS, no general conclusion can be drawn regarding the difference in the degree of diversification between USO and ME. As is reported in (9) there exists a critical value for ρ, denoted by $\tilde{\rho}$, which determines the difference between n_e and n_u. Furthermore, $\tilde{\rho}$ depends negatively on both fixed costs (a) and the share parameter (γ).

An important early application of the DS model to international trade issues is Krugman (1990). Application of our solution method reverses some of the conclusions drawn by Krugman (1980), thereby resolving an anomaly in that model. The Krugman model is a special case of the DS model with $s(q) = \gamma = 1$. The solution of this model can be obtained by applying the results in Proposition 2. In Krugman's model, I is interpreted as the size of the economy. The emergence of international trade leads to an increase in the size of I, which in turn influences the price level, output, and the degree of diversification. Using the results from part a of Proposition 2, it is straightforward to verify the following comparative-static results:

$$(11)\quad \frac{\partial p_e}{\partial I} < 0 \qquad \frac{\partial x_e}{\partial I} > 0 \qquad \frac{\partial n_e}{\partial I} > 0.$$

Krugman (1980) himself uses the DS approach to solve his model. He makes the sharp prediction that the price and output are independent of the size of the economy; that is, an opportunity for international trade has no effect on the wage price of goods or on the output level of each good. In contrast, x_e and p_e as obtained from part a of Proposition 2 do depend on the size of the economy with sign given in (11), and therefore on the opening up of international trade.

Krugman (1980 p. 953) expresses concern about the counterintuitive feature of his solution but claims that we need a more complicated model to sort out the effects of international trade on prices and output levels of each good. Using our solution method, no further complications to the model are needed in order to obtain plausible results. Of course, if I and n are large, as Krugman assumes, then the comparative-static predictions derived from Krugman's solution are good proxies for the correct solutions given by our approach.

II. Increasing Returns to Scale and Isoelastic Cost Functions

In the original DS model, increasing returns to scale exist because of the existence of fixed cost. We now adapt the model in the DS paper and specify the homogeneous production functions for x_i and x_0 as follows:

$$(12)\quad x_i = \ell_i^{1/\beta} \qquad \beta < 1 \qquad x_0 = \ell_0$$

Here ℓ_i denotes the quantity of labor employed to produce good i $(i = 0, 1, \ldots, n)$, and $\beta < 1$ characterizes globally increasing returns to scale in producing good i. Each consumer is endowed with one unit of labor, and the size of the labor force is L. Labor is assumed to be the numeraire and the nominal wage rate is given as $w = 1$. Proposition 3 summarizes the results using our approach.

PROPOSITION 3: *(a) The market equilibrium (ME) exists if $\rho > \beta$. The solutions for*

$x_i = x$, $p_i = p$, x_0, p_0, *and* n *are:*

$$(13)\quad p_e = \left[\frac{\rho(1-\beta)}{\gamma(\rho-\beta)L}\right]^{(1-\beta)/\beta} \qquad p_0 = 1$$

$$(14)\quad x_e = \left[\frac{\gamma L(\rho-\beta)}{\rho(1-\beta)}\right]^{1/\beta} \qquad x_0 = (1-\gamma)L$$

$$(15)\quad n_e = \frac{\rho(1-\beta)}{\rho-\beta} \qquad \varepsilon_i = \frac{1}{\beta-1}$$

$$\varepsilon_{ij} = \frac{\rho-\beta}{(1-\beta)(1-\rho)}.$$

(b) The market equilibrium does not exist if $\rho \le \beta$.

In contrast, the solution obtained by using the DS approach does not exist for this model except if $\rho = \beta$, but at that point the number of firms (and the degree of diversification) is indeterminate. Although $\rho = \beta$ is not impossible, the DS solution method is incorrect in this case. Part a of Proposition 3 shows that a well-defined market equilibrium does exist for values of ρ in excess of β. The DS method finds an equilibrium even though the correct equilibrium does not exist, if $\beta = \rho$.

The intuition behind the "knife-edge" result is quite straightforward. The ratio of marginal cost (MC) to average cost (AC) is constant (i.e., $MC = \beta AC$). For the DS method we must simultaneously satisfy MC = MR, $P = AC$, and $MR = \rho P$, where MR represents marginal revenue and P is the price. This can only be the case if $\beta = \rho$. At that point there is nothing determining n. In our solution method this knife-edge problem disappears because the elasticity of demand depends on n, and the solution satisfies MC = MR, and $P = AC$ as before, but $MR = [\rho(n-1)/(n-\rho)]P$. This can be solved for meaningful solutions if $\beta < \rho$.[7]

[7]This result is not specific to the simple production structure used in (12). In fact, the result holds for any multi-input homogeneous production function provided all factors are variable, so that the cost function is isoelastic.

The knife-edge problem associated with the DS approach is not without practical relevance. Indeed, of the four production functions most commonly used by economists (i.e., Cobb-Douglas, constant elasticity of substitution, Leontief, and linear), only the linear production function is compatible with the DS approach, provided fixed costs exist.[8]

III. Concluding Remarks

We have shown that a simple alternative solution method exists for the DS model—one which makes full use of the structure of that model. We have traced through the implications of this new solution method on the qualitative conclusions derived from the DS model itself and various models derived from it. The implications of our solution method in some cases reverse the conclusions derived on the basis of the DS solution method.

The DS solution method is valid if the number of commodities (the degree of diversification) is large: the validity of the assumption of monopolistic competition depends on a large number of goods. The number of goods, however, is endogenously determined within the model and depends on specifications of technology and endow-

[8]Ethier (1982) applies the ideas of DS on the production side of his model. Since he employs the DS solution method, this model cannot generate a well-defined equilibrium if the production function for producer goods is isoelastic, while our approach can.

Our approach also has relatively minor consequences for the recent dynamic model of Grossman and Helpman (1989). Specifically, it can be shown, using our approach, that the $\dot{n} = 0$ line is nonlinear and downward-sloping and not horizontal as Grossman and Helpman (1989 p. 1268) suggest. This modification is of relatively minor importance for their linearized steady-state analysis. It may, however, gain importance in a full nonlinear dynamical analysis of their model.

The knife-edge result (plaguing the DS and Ethier models) is not a problem in Grossman and Helpman's (1989) model because positive profits exist due to the predetermined nature of n, the number of product varieties. Hence, the constraint $P = AC$ is not operative.

ments. Therefore, one cannot tell a priori whether the number of goods is large, whether the cross price elasticity is trivial, and whether the DS approach gives acceptable approximate solutions unless one knows the explicit specification of the model and has solved the model using our approach.

Our solution method widens significantly the applicability of the DS model. Specifically, whereas the DS solution approach cannot accommodate isoelastic (decreasing) cost functions, our method can. Hence, Cobb-Douglas, CES (constant elasticity of substitution), and Leontief production technologies are not problematic in the DS model, provided our solution method is employed.

A final argument in favor of our approach is the fact that no additional complexity is introduced. The adjustment to the pricing rule is simple and intuitive, and explicit solutions for the endogenous variables in the model can still be calculated easily. On the basis of these arguments we suggest replacing the DS solution method with ours.

REFERENCES

Blanchard, Olivier Jean, and Kiyotaki, Nobuhiro, "Monopolistic Competition and the Effects of Aggregate Demand," *American Economic Review*, September 1987, *77*, 647–66.

Dixit, Avinash and Stiglitz, Joseph E., "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, *67*, 297–308.

Ethier, Wilfred J., "National and International Returns to Scale in the Modern Theory of International Trade," *American Economic Review*, June 1982, *72*, 389–405.

Grossman, Gene M. and Helpman, Elhanan, "Product Development and International Trade," *Journal of Political Economy*, December 1989, *97*, 1261–83.

_____ **and** _____, "Comparative Advantage and Long-Run Growth," *American Economic Review*, September 1990, *80*, 796–815.

Krugman, Paul R., "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review*, December 1980, *70*, 950–9.

Romer, Paul M., "Endogenous Technological Change," *Journal of Political Economy*, October 1990, *98*, S71–S102.

Rotemberg, Julio J., "The New Keynesian Microfoundations," in Stanley Fischer, ed., *NBER Macroeconomics Annual, 1987*, Cambridge, MA: MIT Press, 1987.

Tirole, Jean, *The Theory of Industrial Organization*, Cambridge, MA: MIT Press, 1988.

Monopolistic Competition and Optimum Product Diversity: Reply

By AVINASH K. DIXIT AND JOSEPH E. STIGLITZ*

Our model of monopolistic competition and product diversity (Dixit and Stiglitz, 1977) was intended for the large-group case. Therefore we used an approximation neglecting from the own and cross elasticities terms of order $1/n$, where n is the equilibrium number of firms. Xiaokai Yang and Ben J. Heijdra (1993) have generalized the solution for a special case of our model, retaining some of these terms but assuming a unitary elasticity of substitution between the monopolistic group and the numeraire good. We will discuss their model by placing it in an even more general framework.

The underlying monopolistically competitive equilibrium of a symmetric group is governed by two conditions: each firm's maximization equates marginal revenue and marginal cost, and free entry equates average revenue and average cost. Let n denote the number of products or firms in the group, x the output of each such firm, and p the price of each product in the group relative to the competitive numeraire. Let $C(x)$ be each firm's cost of production, and let $\varepsilon(p,n)$ be the elasticity of demand perceived by each firm. Then the two conditions can be written

$$(1) \qquad p\left[1-\frac{1}{\varepsilon(p,n)}\right]=C'(x)$$

$$(2) \qquad p=C(x)/x.$$

Further, let the demand relation in equilibrium be

$$(3) \qquad p=D(x,n).$$

These three equations determine p, x, and n. Unless one wants a closed-form solution, the various functions, including the elasticity $\varepsilon(p,n)$, can have any functional form (restricted only by minor requirements for existence and the relevant second-order conditions). This formulation subsumes not only Yang and Heijdra's model, but also other widely different models including the one based on Hotelling-like spatial product differentiation. Indeed, Elhanan Helpman and Paul Krugman (1985 Chapters 6–8 [especially section 7.1]) develop the monopolistic competition model for their international trade applications in just such a general way.

Our original (1977) model started with a CES (constant elasticity of substitution) utility function and then employed an approximation ignoring some terms of order $1/n$ to make ε constant. Yang and Heijdra dispense with this approximation, using instead

$$(4) \qquad \varepsilon(p,n)=\sigma-\frac{\sigma-1+\theta(n^{1/\rho}p)}{n}$$

where σ, required to be greater than 1, is the elasticity of substitution between any pair of goods within the monopolistically competitive group, $\rho=(\sigma-1)/\sigma$. Finally, θ is the share of the consumer's expenditure on the monopolistically competitive group and is a function of the group price index $(n^{1/\rho}p)$. Since Yang and Heijdra insist on a closed-form solution, they have to assume that θ is constant, which amounts to requiring a unitary elasticity of substitution in utility between the monopolistically competitive group and the perfectly competitive numeraire good.

We believe that Yang and Heijdra's misplaced emphasis on closed-form solutions

*Department of Economics, Princeton University, Princeton, NJ 08544, and Department of Economics, Stanford University, Stanford, CA 94305, respectively.

has led them astray. The desirable features they claim to have added can be attained in far better and more general ways in the formulation (1)–(3) above. To the extent that closed forms are needed for simple applications in macroeconomics and international trade, for example, Yang and Heijdra's formulation adds a minor wrinkle at one point but incurs a large cost in reduced generality elsewhere.

I. Justification for the Approximation

Yang and Heijdra (1993) argue that since the equilibrium number n_e of firms is endogenous, an assumption neglecting $1/n_e$ needs to be justified in terms of the underlying parameters. This is true, but not at all hard to do. If one insists on a closed form, one must resort to the unitary-elasticity case used by Yang and Heijdra. There, n_e is large if and only if

$$\gamma I(1-\rho)/a \gg 1 \tag{5}$$

where γ is the share of the monopolistic group in expenditure, I is the economy's labor endowment, ρ measures substitution within the monopolistic group, and a is the fixed cost of producing each variety in this group. However, without a closed-form solution, we can get a qualitative idea of which parameter combinations are conducive to a large n_e by examining the total differential of the system (1)–(3) above. In rough terms, we want low fixed costs and imperfect substitution between products in the group.

We should also point out that Yang and Heijdra destroy the force of their own criticism when they assume that cross elasticities are negligible. Here they see no need to express the condition in terms of exogenous parameters: "in any application of this model one must ascertain *ex post* that ε_{ij} is indeed close to zero *in equilibrium*" (Yang and Heijdra, 1993 p. 296 [italics in the original]).

II. Internal Consistency of the Approximation

Having decided to treat $1/n$ as small, we do so with consistency throughout our paper. Yang and Heijdra (1993) emphasize the terms of order $1/n$ in the own elasticities ε_i but ignore them in the cross elasticities ε_{ij}. This is particularly strange in view of the fact that the error terms in treating $\varepsilon_i \approx -\sigma$ and $\varepsilon_{ij} \approx 0$ are *exactly the same*, namely, $[\sigma - 1 + \theta(q)]/n$. It does not seem logically consistent to emphasize the error in one place and to ignore it in another in the same model.

If Yang and Heijdra want to consider a group sufficiently small that $1/n$ is not negligible, they should do so consistently. When ε_{ij} is significantly different from zero, it raises serious doubts concerning the validity of the price–Nash game stipulated in the model, and alternatives including tacit collusion and entry deterrence should be explored.

III. Passage to a Competitive Limit

Yang and Heijdra (1993) argue that there is merit in letting elasticities increase with n. This is true in some applications, although in others (e.g., Helpman and Krugman, 1985 Chapters 7–8) it makes no difference to the qualitative results. In any case, a correction of order $1/n$ does not capture such dependence in a satisfactory way. In Yang and Heijdra's equation [see (4), above], as n goes to infinity the own elasticity will rise in numerical value to σ. If one wants a perfectly competitive limit of the model, as many would, the own and cross elasticities will have to rise to *infinity*, not to a finite limit.

IV. Optimality of the Equilibrium

While the limiting case of perfect competition and full optimality may be of some interest, we found the equilibrium based on our approximation to be second-best (break-even constrained) optimal. This result is sensitive to the specification, as we pointed out. What is really needed is a constant ratio of average and marginal utilities (or consumer surplus and revenue), and a constant elasticity of demand allows that. However, Yang and Heijdra's (1993) modification can only shift the equilibrium in a

small way when n is large; the numbers and prices will be close to those in our approximation, and therefore their equilibrium will be close to being constrained optimal. Serious objections to the optimality result must again be based on more substantial departures from our specification. For a spatial-differentiation example in which the dimension of the attribute space makes a difference, see Stiglitz (1986).

V. Existence of Equilibrium

Yang and Heijdra (1993) point out that if ε is constant in (1) and increasing returns take the power form, the system (1)–(3) is in general inconsistent. Write the cost function as $C(x) = x^{\beta}$ with $\beta < 1$; then $C'(x) = \beta C(x)/x$, and dividing (1) by (2) gives

$$1 - 1/\varepsilon = \beta$$

which is generally impossible when ε and β are separately and exogenously specified. In Dixit and Stiglitz (1977) the cost function had a fixed component with constant marginal cost, so the problem did not arise. Yang and Heijdra propose to rescue matters for the power cost function by letting ε vary endogenously as in (4) above. However, for large n the variation of ε is very slow, and the equilibrium shifts dramatically when some parameter changes by a small amount. The problem of such "fragile" equilibria is in practice not much better than nonexistence (see e.g., Lawrence Summers, 1991). Again we think that Yang and Heijdra's aim is much better accomplished in the general setting of (1)–(3) above.

VI. Price Paid for Abandoning the Approximation

To introduce a term of order $1/n$ Yang and Heijdra (1993) have restricted the model to the case of a unitary elasticity of substitution between the monopolistic group and the numeraire good. We judge the trade-off to be quite unfavorable; the drastic reduction in parameter space of the model buys a correction that in practice is of minor significance.

This case has been used in some applications to make some broad thematic points about the nature of trade or the possibility of underemployment equilibrium, where the exact value of the elasticity was irrelevant. However, the presence or absence of $1/n$ terms in elasticities was equally irrelevant there.

Finally, we should note that Yang and Heijdra are not the first to add the $1/n$ terms to our model, although they appear to have been the first to be published doing so. To record priority, we should mention an unpublished note by Kelvin Lancaster (1980); doubtless there are others unknown to us.

REFERENCES

Dixit, Avinash K. and Stiglitz, Joseph E., "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, *67*, 297–308.

Helpman, Elhanan and Krugman, Paul, *Market Structure and Foreign Trade*, Cambridge, MA: MIT Press, 1985.

Lancaster, Kelvin, "Pettengill versus Dixit-Stiglitz: A Third-Party Intervention," working paper, Columbia University, 1980.

Stiglitz, Joseph E., "Towards a More General Theory of Monopolistic Competition," in Maurice Peston and Richard Quandt, eds., *Prices, Competition and Equilibrium*, Oxford: Philip Allan, 1986, pp. 22–69.

Summers, Lawrence, "Comment" in Olivier Blanchard and Stanley Fischer, eds., *NBER Macroeconomics Annual 1991*, Cambridge, MA: MIT Press, 1991, pp. 54–9.

Yang, Xiaokai and Heijdra, Ben J., "Monopolistic Competition and Optimum Product Diversity: Comment," *American Economic Review*, March 1993, *83*, 295–301.

Should Governments Learn to Live with Inflation? Comment

By JONAS AGELL AND BENGT-CHRISTER YSANDER*

Economists have for a long time tried to convince suspicious policymakers of the merits of various indexation policies as a way of insulating the economy from the potentially pervasive effects of inflation. In an interesting recent paper, Stanley Fischer and Lawrence Summers (1989; hereafter FS) argue that the suspicion is warranted. Relying on the dynamic consistency theories of inflation developed by Finn Kydland and Edward Prescott (1977) and Robert Barro and David Gordon (1983), they reach two primary conclusions:

> First, recognizing that inflation is a choice variable that will be affected by changes in nominal institutions substantially undercuts the case for inflation mitigation measures by governments that have not established a firm anti-inflationary reputation. It is easy to construct examples in which the costs of the extra inflation that results from inflation palliation outweigh the direct benefits of the lower cost of a given inflation rate. Second, foregoing indexation is to some extent a substitute for developing a reputation for pursuing anti-inflation policies. Nations with strong anti-inflationary reputations can more easily afford indexation policies than nations without such reputations. [1989 p. 382]

The generality of the claims of FS fits less well with the specificity of their theoretical model. First, the FS analysis is appropriate for economies in which wage-setting takes place in a decentralized fashion and, as a corollary, wage-setters do not behave strategically. Yet they offer their model as an explanation for the fact that "...most industrial countries do relatively little to mitigate its [inflation's] adverse effects" (FS, 1989 p. 382). We show below that their conclusions need not apply to economies in which wage-setting is managed in a more centralized and strategic manner. Since a number of industrialized economies fall into this latter category or between the two extremes of centralized and decentralized wage-setting, there is little ground for FS's general presumption against inflation palliation.

Second, in the setup of FS the only way for unanticipated inflation to affect the economy is through a standard expectations-augmented Phillips curve. Yet, probably much more important effects are the intersectoral and intergenerational redistributions of income and wealth associated with unanticipated inflation. These redistribution effects provide a simple zero-sum rationalization of why economies with unindexed institutions may end up in a high-inflation equilibrium.

I. Review of the Fischer-Summers Analysis

Following the model of Barro and Gordon (1983) of the credibility problem of low-inflation monetary policy, FS assume a short-run Phillips curve

$$(1) \qquad U = U_n - a(p - p^e)$$

where U is the unemployment rate, U_n is the natural rate of unemployment, p is the log of the price level, and p^e is the log of the expected price level. Adopting the convention that the initial price level is unity (i.e., $p_0 = 0$), the inflation rate simply becomes p.

*Professor Bengt-Christer Ysander, of Uppsala University and the Royal Swedish Academy of Sciences, passed away on March 29, 1992. Jonas Agell is at the Institute for International Economic Studies, S-10691 Stockholm, Sweden. The authors appreciate the comments of Susanne Ackum, Anders Forslund, Bertil Holmlund, Torsten Persson, and in particular, an anonymous referee.

The government sets p to minimize the loss function

$$L = \tfrac{1}{2}U^2 + \tfrac{1}{2}bp^2 \tag{2}$$

where the social efficiency costs of actual (as opposed to expected or unanticipated) inflation are measured in an Harberger-like manner by the quadratic term $\frac{1}{2}bp^2$.[1] In the one-shot game between a myopic government and the private sector considered by FS, equilibrium inflation (in the sense that $p = p^e$) becomes

$$p = (a/b)U_n. \tag{3}$$

Equation (3) lies at the heart of the policy conclusions of FS. Equilibrium inflation increases with a and decreases with b. Interpreting these results in terms of indexation, FS conclude that wage indexation (which reduces a and makes the Phillips curve steeper) lowers inflation and increases social welfare. More provocatively, they also contend that most other forms of indexation increase inflation and decrease social welfare by decreasing b.[2]

II. Indexation in a Model of Strategic Union–Government Interaction

Let us start with the case of strategic wage-setting in an economy with unindexed institutions. Following John Driffill (1985), Henrik Horn and Torsten Persson (1988), and others, we consider a game between an all-encompassing monopoly union setting the rate of wage inflation and a government controlling the rate of general inflation.

While the government's objective function (2) carries over to our model, describing union behavior requires some additional notation and definitions. We assume that the union cares about the real wage after taxes (Tor Hersoug, 1984). To capture the well-known effect of nonindexation of tax brackets in a progressive income-tax system, we consider a real disposable income function of the form

$$Y = \frac{W^\tau}{P} \tag{4}$$

where $0 \le \tau \le 1$ and capital letters denote wage and price levels. The parameter τ represents the degree of tax progression. The smaller is τ, the more substantial is the bracket-creep induced by higher wage inflation.

Unemployment depends positively on the log of the real wage:

$$U = w - p. \tag{5}$$

The objective of the union is to maximize

$$-\tfrac{1}{2}U^2 + a_1(\tau w - p) - \delta\tfrac{1}{2}bp^2. \tag{6}$$

The first and second terms are of a standard form. As in Driffill (1985), the utility function is quadratic in unemployment and linear in the real wage after tax. We also allow the union to care about actual inflation; an all-encompassing union should account for the fact that its members ultimately bear some of the efficiency costs of inflation (Richard Jackman, 1988). These efficiency costs are reflected by the third term. The constant δ $(0 \le \delta \le 1)$ represents the share of the social efficiency costs of inflation that is internalized by the union. When $\delta = 1$, government and union give the same weight to the costs of actual inflation.

Consider a single-stage Stackelberg game, in which the union (the leader) sets the wage before the government determines the price level. Apart from conforming with FS's benchmark case of a myopic accommodating government, this way of structuring the game also seems to parallel the institutional setup in many European countries (see e.g., Hersoug, 1985; Horn and Persson, 1988). Minimizing (2) subject to (5), treating w as parametrically given, gives us the govern-

[1] To facilitate comparison with our model of strategic union–government interaction, we assume that the parameter k in equation (2) of FS, representing the government's target rate of unemployment, is zero.

[2] Evaluating (2) in a fulfilled-expectations equilibrium, it is easy to see that the government's loss function is decreasing in b.

ment's reaction function:

$$(7) \qquad p = \frac{w}{1+b}.$$

We assume that the union knows the government's reaction function. Maximizing (6) subject to (5) and (7) yields the optimal wage:

$$(8) \quad w^* = (1+b)a_1[\tau(1+b)-1] \times [b(\delta+b)]^{-1}$$

which combined with (7) gives equilibrium inflation as

$$(9) \quad p^* = a_1[\tau(1+b)-1][b(\delta+b)]^{-1}.$$

In the model of FS, equilibrium unemployment, U_n, is taken as given and unaffected by changes in nominal institutions. In our model with wage-setting agents, equilibrium unemployment is a choice variable conditioned on the institutions of the economy. Combining (5), (8), and (9), we obtain

$$(10) \quad U^* = a_1[\tau(1+b)-1](\delta+b)^{-1}.$$

In the verbiage of the literature, equations (9) and (10) characterize the "discretionary," or "time-consistent," solution to our stabilization policy game.[3] While both the government and the union would prefer lower inflation, neither of them has an incentive to deviate from the wage and price decisions given in (8) and (9). However, a cursory inspection reveals that the equilibrium depends on the institutional environment. By exercising different indexation options, both the government and the union can alter the policy framework which both players take as given when setting prices and wages.

The main conclusion of FS is that all measures that decrease the efficiency costs of actual inflation (like paying interest on nominal money holdings or reducing the deadweight losses induced by nominalistic capital taxation by indexing the tax base) are likely to be counterproductive. This result need not hold in economies where wage-setters behave strategically.

Differentiating (9) and (10) with respect to b, we find that

$$(11) \quad \frac{\partial p^*}{\partial b} = a_1\{(\delta+b)(1-\tau) - b[\tau(1+b)-1]\}[b(\delta+b)]^{-2}$$

$$(12) \quad \frac{\partial U^*}{\partial b} = a_1[1-\tau(1-\delta)](\delta+b)^{-2}.$$

Only in the special case when $\tau = 1$ and $\delta = 0$ do we obtain results analogous to those of FS. Indexation measures that reduce b then lead to higher inflation, while unemployment remains unchanged. Consider next the case when $\tau = 1$ and when the union internalizes some of the excess burden of inflation (i.e., $\delta > 0$). From (11), we see that a lower value of b again leads to higher inflation. However, from (12), it now follows that unemployment decreases. As the union knows the government's reaction function (7), it also realizes that a lower value of b implies a more inflationary policy rule. Consequently, a union that cares about inflation moderates its wage claims, which in turn implies lower real wages and unemployment. An analysis of the net welfare effects of decreasing b must set the gains from lower unemployment against the costs from higher inflation.[4]

Even when the union does not care about inflation, it is easy to find examples in which protection against the efficiency costs of inflation is beneficial. When $\tau < 1$, we note from (12) that a decrease in b is associated with a decrease in unemployment irrespective of the value of δ. The bracket-creep

[3] In the following, we assume that the tax system is not "too" progressive, in the sense that $\tau \geq 1/(1+b)$. As a consequence, p^* and U^* are always nonnegative. The intuition is discussed in what follows.

[4] Using (2), we can show that the net outcome of decreasing b is higher welfare whenever $\delta = 1$.

implied by a progressive and unindexed tax on earned income means that the union again moderates its wage claims when a lower value of b makes the government less prone to inflation-fighting. This mechanism also explains why $\partial p^*/\partial b$ is an increasing function of τ. Indeed, if τ is sufficiently small, a lower value of b may reduce both unemployment and inflation.

While the preceding argument casts doubt about FS's general claims concerning the effects of mitigating the efficiency costs of inflation, it also suggests that nonindexed tax brackets might be a lesser evil than generally perceived. From (9) and (10), it is easy to see that we can reach a social optimum with both inflation and unemployment equal to zero by setting the degree of tax progression as

$$\tau^* = \frac{1}{1+b}. \tag{13}$$

The inflation and unemployment characterizing the initial equilibrium is due to the fact that the union, unlike the government, wants to trade lower employment for higher real wages. The trade-off facing the union depends, however, on the degree of tax progression. When (13) holds, the union realizes that bracket-creep and inflation will undercut any increase in nominal wages. As the union still dislikes unemployment, the net outcome is a realignment of government and union objectives, which in turn implies an equilibrium with zero inflation and unemployment. As a corollary, it is easy to show that indexation of tax brackets is unwarranted. By removing one of the costs that wage inflation imposes on the union, indexation of tax brackets produces more inflation and unemployment.[5]

[5]To show this, replace equation (4) by the disposable income function $Y = (W/P)^{\tau}$ and derive the corresponding expressions for inflation and unemployment.

III. Some Alternative Games

Game-theoretical policy analysis seldom generates robust insights. Our union–government game is no exception. Abandoning the notion of a strong union and a weak government, our results will change. In a Nash equilibrium, the government and the union simultaneously set prices (government) and wages (union), while taking each other's optimal reactions as given.[6] The union's reaction function, obtained by maximizing (6) subject to (5), holding p constant, is given by

$$w = p + a_1\tau. \tag{14}$$

Combining (14) and (7), equilibrium inflation and unemployment become

$$p^* = a_1\tau/b \tag{15}$$

$$U^* = a_1\tau. \tag{16}$$

The Nash equilibrium of our model is closely related to the discretionary equilibrium of the model of FS. In both cases, private agents act nonstrategically vis-à-vis the government. We may thus expect results in line with those of FS. This is indeed the case. Equations (15) and (16) imply that indexations that decrease the efficiency costs of inflation increase equilibrium inflation while leaving unemployment unchanged. As the union treats inflation as beyond its influence, it has no incentive to moderate wages in response to a smaller value of b. For the same reason, it is also easy to show that inflation protection of tax brackets has no effects on inflation and unemployment.

Consider next a Stackelberg game in which the government is the leader. From

[6]We note in passing that this setup can be interpreted as a game between an accommodating government and an aggregate of several unions, where the individual union is small enough to ignore the government's reaction to its wage decision (see e.g., Andrew Oswald, 1979).

the union's reaction function (14), it is clear that equilibrium unemployment is the same as unemployment in the Nash solution. Equilibrium inflation, on the other hand, is zero, the reason being that the government now uses its influence on union wage formation to induce the union to obtain its real wage target at a lower nominal wage than in the Nash equilibrium.

That a strong government may produce a better outcome than a weak accommodating one is hardly surprising. This result is, however, not robust. If the government cares about its tax revenue, the nonindexation of tax brackets provides an incentive for generating extra inflation. From the union's reaction function, the government knows that higher inflation increases wages proportionally and that the resulting bracket-creep increases real tax revenue whenever $\tau < 1$. In contrast to the case of a strong union and weak accommodating government discussed in the preceding section, inflation protection of tax brackets now leads to lower inflation.

IV. A Zero-Sum Model of Inflation

In their systematic account of the real effects of inflation, Fischer and Franco Modigliani (1978 p. 827) conclude that "...the wealth redistributions arising from unanticipated inflation are large, of the order of 1 percent of GNP per 1 percent unanticipated increase in the price level." While it is difficult to be more precise about the direction of these redistributions, two broad tendencies stand out. First, in an unindexed economy, unexpected inflation is likely to imply redistributions from the old to the young, partly because the ratio of net nominal to real assets typically rises with age and partly because inflation erodes the value of unindexed social security (Fischer and Modigliani, 1978 pp. 823–6). Second, in countries where the government is a net debtor, it benefits from surprise inflation.

These wealth transfers may amount to much more than mere bookkeeping. Consider the effects of adding a retired generation to the model of union–government interaction of Section II. This third player holds nominally denominated government debt and obtains unindexed social security but has no influence on the price and wage decisions of government and union. Both government and union then gain from unanticipated inflation: the government by getting its debt amortized, union members by an induced transfer from the retired generation. Formally, the government minimizes the loss function

$$(2') \qquad \tfrac{1}{2}U^2 + \tfrac{1}{2}bp^2 - \gamma_g(p - p^e)$$

while the union maximizes

$$(6') \qquad -\tfrac{1}{2}U^2 + a_1(\tau w - p) - \delta\tfrac{1}{2}bp^2 + \gamma_u(p - p^e).$$

The last term in each expression represents the distributional gain from unexpected inflation.[7] While p is the actual price level, p^e is the price level expected by the retired generation. We assume that both union and government treat p^e as exogenously given (in a fulfilled-expectations equilibrium, p^e equals the equilibrium inflation rate). The parameters γ_u and γ_g ($\gamma_g, \gamma_u > 0$) represent the strengths of the union's and the government's incentives to create unexpected inflation.

Consider again the Stackelberg game of Section II, in which the union (leader) sets the wage before the government (follower) determines the price level. Repeating the steps followed above, we obtain

$$(9') \quad p^* = \{b\gamma_g + \gamma_u + a_1[\tau(1+b) - 1]\} \times [b(\delta + b)]^{-1}$$

$$(10') \quad U^* = \{\gamma_u - \gamma_g\delta + a_1[\tau(1+b) - 1]\} \times (\delta + b)^{-1}.$$

[7] Allowing for quadratic terms in unexpected inflation would complicate the algebra, without affecting the essence of our results.

Differentiating, we find that

$$\text{(17a)} \qquad \frac{\partial p^*}{\partial \gamma_g} > 0$$

$$\text{(17b)} \qquad \frac{\partial U^*}{\partial \gamma_g} < 0 \qquad \text{if } \delta > 0$$

$$\text{(17c)} \qquad \frac{\partial p^*}{\partial \gamma_u} > 0$$

$$\text{(17d)} \qquad \frac{\partial U^*}{\partial \gamma_u} > 0.$$

The intuition is not hard to find. Increasing the gains from unexpected inflation increases equilibrium inflation. In the case of the union, the incentive to create inflation operates in a roundabout way. When γ_u increases, the union sets a higher nominal wage, knowing that the accommodating government thereby is induced to create extra (unexpected) inflation. This mechanism also explains why an increase in γ_u leads to higher unemployment. As the government only partially accommodates the wage push, real wages and unemployment increase. Finally, we note that an increase in γ_g reduces unemployment, at least as long as the union internalizes some of the deadweight burden of additional inflation. When the government is given a marginal incentive to create surprise inflation, a union that cares about inflation will moderate its wage claims, which in turn implies lower unemployment and real wages.

From the above, it is clear that reducing the scope of intergenerational redistributions from the retired generation to the collective of younger trade-union members unambiguously increases social welfare. By indexing social security and private loan contracts, the value of γ_u declines, which lowers inflation and unemployment. Also, with a social loss function quadratic in U and p, it is easy to show that measures that reduce γ_g (e.g., indexation of government debt instruments) increase welfare, since the gain in the form of lower inflation then always dominates the cost in terms of higher unemployment.

The policy lesson is straightforward. For every loser from unanticipated inflation there is a gainer. If the gainers also happen to be key players in the inflationary process, the outcome is high inflation. This zero-sum rationalization of inflation also goes some way toward explaining the common practice of indexing social-security benefits.[8] However, it also raises the question of why other forms of protection against unexpected wealth redistributions are so rarely seen. In particular, by eliminating the other main channel for wealth redistributions from unexpected inflation, widespread indexation of private and government loan contracts may reduce inflation substantially.

V. Conclusions

There is no general time-consistency argument against indexation policies. In economies where wage-setters behave strategically, governments may have much to gain from better inflation protection. If we accept the view that much of the driving force in the inflationary process comes from the zero-sum redistributions from unexpected inflation, we also have a powerful argument for indexations of social security and private and government loan contracts.

REFERENCES

Barro, Robert J. and Gordon, David B., "A Positive Theory of Monetary Policy in a Natural Rate Model," *Journal of Political Economy*, August 1983, *91*, 589–610.

Driffill, John, "Macroeconomic Stabilization Policy and Trade Union Behaviour as a Repeated Game," *Scandinavian Journal of Economics*, 1985, *87* (2), 300–26.

[8]The conclusion of FS that "measures that protect Social Security Benefits from being eroded by increased inflation...may significantly increase equilibrium inflation rates and reduce welfare by eliminating important sources of political opposition to inflation" (FS, 1989 p. 387) seems to rely on a somewhat sanguine view of the political process.

Fischer, Stanley and Modigliani, Franco, "Towards an Understanding of the Real Effects and Costs of Inflation," *Weltwirtschaftliches Archiv*, 1978, *114* (4), 811–33.

_____ **and Summers, Lawrence H.,** "Should Governments Learn to Live with Inflation?" *American Economic Review*, May 1989 (*Papers and Proceedings*), *79*, 382–7.

Hersoug, Tor, "Union Wage Response to Tax Changes," *Oxford Economic Papers*, March 1984, *36*, 37–51.

_____, "Workers versus Government—Who Adjusts to Whom?" *Scandinavian Journal of Economics*, 1985, *87* (2), 270–92.

Horn, Henrik and Persson, Torsten, "Exchange Rate Policy, Wage Formation and Credibility," *European Economic Review*, October 1988, *32*, 1621–36.

Jackman, Richard, "Wage Formation in the Nordic Countries Viewed from an International Perspective," Seminar Paper No. 415, Institute for International Economic Studies, Stockholm, 1988.

Kydland, Finn E. and Prescott, Edward C., "Rules Rather Than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

Oswald, Andrew J., "Wage Determination in an Economy with Many Trade Unions," *Oxford Economic Papers*, November 1979, *31*, 369–85.

Should Governments Learn To Live with Inflation? Reply

By STANLEY FISCHER AND LAWRENCE SUMMERS*

Until recently, it was common for economists to urge the widespread adoption of indexation, and for policymakers in low-inflation economies to reject that advice as a counsel of despair, likely only to increase inflation and reduce economic welfare. The main contribution of our 1989 paper "Should Governments Learn to Live with Inflation?" was to provide a coherent model that explains the policymakers' viewpoint. As we stated, "It is easy to construct examples in which the costs of the extra inflation that results from inflation palliation outweigh the direct benefits of the lower cost of a given inflation rate" (p. 382).

While we were pleased to have found an explanation for a hitherto puzzling conflict of views between economists and policymakers, no reader of our article should have come away from it with the view that we either argued or believe that inflation palliation always reduces economic welfare. We presented examples where that is not the case and provided as thorough a discussion as space permitted of why we believe that indexation is useful in high-inflation economies.

Jonas Agell and Bengt-Christer Ysander (1993) appear to read our paper incorrectly as arguing that inflation palliation always or almost always reduces welfare. They present two models in which inflation palliation improves welfare, as it does in some of the models in our paper. What should the reader make of their models and results? The first model is one in which wages are set strategically by a union that faces a nonindexed tax system and that also cares about inflation. This is a standard model of strategic wage-setting, which has been used by John Driffill (1985) and Henrik Horn and Torsten Persson (1988) among others. It remains relevant to some European countries.

In this model, the impact of inflation mitigation on welfare depends on parameter values, as it does in our 1989 paper when inflation is uncertain. In Agell and Ysander's model, by setting the progressiveness of wage taxation appropriately, it is even possible to reach the social optimum in their nonprecommitment wage- and price-setting game. In this situation, the union, which is penalized for inflation by the nonindexation of income tax, internalizes some of the social costs of inflation.

Agell and Ysander's model produces a result that is very much in the spirit of the basic result in our paper, for as they show, it is optimal *not* to index wage taxation when income tax progression is set so that the first-best outcome is attained. This is another example of a model in which the introduction of indexation may reduce social welfare—though to repeat, there are parameter values for which other inflation-mitigation measures improve social welfare.

Agell and Ysander appear to place more weight on the final model presented in their paper, in which unanticipated inflation redistributes wealth and is taken into account by the government and the wage-setting union. When the price level is higher than expected, both the government and the workers benefit. They gain at the expense of retirees who are not explicitly included in the model. The result here is that inflation mitigation reduces the inflation rate; that is, when the government and the union each benefit less from higher inflation, the inflation rate declines.

*Fischer: Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, and NBER; Summers: The World Bank, 1818 H Street, N.W., Washington, DC 20433, Harvard University, and NBER. We are grateful to Lant Pritchett for assistance.

While this result is correct in the context of the model, we find the objective functions used in the model extremely unpersuasive. U.S. politicians would be astonished to learn that retirees have no political voice or impact on government preferences. We are also surprised that unions, which have extensive pension funds that are damaged by unexpected inflation whether invested in either bonds or stocks, are treated as welcoming unanticipated inflation.

Agell and Ysander describe unexpected inflation as producing zero-sum redistributions in their model. In fact, unexpected inflation benefits all participants whose decisions are registered in their model. We would have expected a loss function for the government that penalizes unanticipated redistributions in general, rather than rewarding redistributions away from the retired. Such an effect could be represented by including a quadratic term in unanticipated inflation in the objective function of the government. The inclusion of such a term does not change the conclusions of nonstochastic versions of the models used in either our 1989 paper or Agell and Ysander's comment.

Furthermore, changes in the cost of unanticipated inflation do not affect the expected inflation rate in a stochastic version of our basic model, as discussed at the beginning of section II of our 1989 paper. Since the expected inflation rate does not change in that model when the costs of unanticipated inflation decline (corresponding to increased indexation) and since policy can do nothing in that model to reduce the variance of the inflation rate, inflation mitigation would increase economic welfare. This is hardly surprising, since there is no feedback between the mitigation of the costs of inflation and the government's choice of the inflation rate.

We find nothing in Agell and Ysander's paper that would cause us to revise the general conclusions presented in the last two paragraphs of our 1989 paper (p. 387):

> This line of argument suggests that certain forms of inflation mitigation such as measures that protect Social Security Benefits from being eroded by increased inflation, or that prevent inflation from causing sharp increases in business tax burdens, may significantly increase equilibrium inflation rates and reduce welfare by eliminating important sources of political opposition to inflation. On the other hand, measures like the use of inflation accounting or the issuance of indexed bonds, that do not undercut important anti-inflation lobbies may have less-pronounced inflation-enhancing effects.
>
> ...The general principle seems to be that in the presence of distortions, policies that would represent Pareto improvements if behavior were unchanged by their implementation, may be undesirable once their incentive effects are recognized.

REFERENCES

Agell, Jonas and Ysander, Bengt-Christer, "Should Governments Learn to Live with Inflation? Comment," *American Economic Review*, March 1993, *83*, 305–11.

Driffill, John, "Macroeconomic Stabilization Policy and Trade Union Behaviour as a Repeated Game," *Scandinavian Journal of Economics*, 1985, *87* (2), 300–26.

Fischer, Stanley and Summers, Lawrence, "Should Governments Learn to Live with Inflation?" *American Economic Review*, May 1989 (*Papers and Proceedings*), *79*, 382–7.

Horn, Henrik and Persson, Torsten, "Exchange Rate Policy, Wage Formation and Credibility," *European Economic Review*, October 1988, *32*, 1621–36.

ERRATUM

Government Size and Economic Growth: A New Framework and Some Evidence from Cross-Section and Time-Series Data: Reply

By RATI RAM*

Please read $\dot{Y}$ in place of $\dot{G}$ on the *left*-hand side of each of the four equations for which regression estimates are reported on pp. 282–3 of my "Reply" in the March 1989 issue of this *Review* (Volume *79*, No. 1, pp. 281–4).

---

* Department of Economics, Illinois State University, Normal, IL 61761-6901.

GREAT EXPECTATIONS

New Economics for a New Administration

JUMPSTART
The Economic Unification of Germany
Gerlinde Sinn and Hans-Werner Sinn

"*Jumpstart* represents the most authoritative and well-reserached analysis of German unification to date. The authors present a lucid and cogent assessment of present and future prospects in East Germany, and their sometimes controversial policy recommendations are based on sound economic reasoning rather than legal of political considerations. German policymakers who ignore the message of this book do so at their own peril." — Michael C. Burda, INSEAD

272 pp. $24.95 (May)

OPENNESS FOR PROSPERITY
Essays in World Economics
Herbert Giersch

Openness for Prosperity brings together Giersch's major essays in macroeconomic policy, written or published over the past two and a half decades.

385 pp. $45.00 (June)

A THEORY OF INCENTIVES IN PROCUREMENT AND REGULATION
Jean-Jacques Laffont and Jean Tirole

" The research reference on incentive regulation."
— Thomas Palfrey, California Institute of Technology

Laffont and Tirole develop a synthetic approach, with a particular, though not exclusive, focus on the regulation of natural monopolies such as military contractors, utility companies, and transportation authorities.

"This book is a major contribution to what the authors aptly call the 'new regulatory economics'. In my opinion, it will guide the economics research on regulation for many years to come."— Michael H. Riordan, Boston University

728 pp., 44 illus.
$37.50 (April)

POSTWAR ECONOMIC RECONSTRUCTION AND LESSONS FOR THE EAST TODAY
edited by Rudiger Dornbusch, Wilhelm Nolling, and Richard Layard

264 pp., 36 illus.
$29.95 (May)

THE ECONOMICS OF MIDDLE EAST PEACE
Views from the Region
edited by Stanley Fischer, Dani Rodrik, and Elias Tuma

This landmark collection of essays addresses the question: Given peace in the Middle East, then what? Economists representing Egypt, Israel, Jordan, Lebanon, and the Occupied Territories offer differing perspectives on what would be required to redirect, rejuvenate, and sustain the economies of the region.

376 pp. $42.50 (May)

ECONOMIC TRANSFORMATION THE MEXICAN WAY
Pedro Aspe

Mexico's current Minister of Finance and Public Credit offers an inside look at attempts to modernize Mexico's economy through the 1970s and 1980s.

Lionel Robbins Lectures
160 pp., 42 illus.
$22.50 (July)

THE ROCKY ROAD TO REFORM
Adjustment, Income Distribution, and Growth in the Developing World
edited by Lance Taylor
These case studies provide valuable insights into the difficulty of establishing answers to the fundamental question of why nations grow at different rates, with inequitable patterns of wealth and income.
500 pp. , 13 illus. $40.00 (July)

POLITICAL ECONOMY OF POLICY REFORM IN DEVELOPING COUNTRIES
Anne O. Krueger
Insight into the economic and political variables that determine the success or failure of economic policy determination and evolution in developing countries.
The Ohlin Lectures 140 pp. $22.50 (June)

COMPETITIVENESS, CONVERGENCE, AND INTERNATIONAL SPECIALIZATION
David Dollar and Edward N. Wolff
The authors look at claims that a deindustrialized United States is on the road to second-rate status in the global marketplace and find them to be both unfounded and simplistic.
232 pp. $35.00 (March)

AGGREGATION
Aggregate Production Functions and Related Topics
Franklin M. Fisher
400 pp. $45.00

OPTIMAL PRICING, INFLATION, AND THE COST OF PRICE ADJUSTMENT
edited by Eytan Sheshinski and Yoram Weiss
These collected articles constitute what is perhaps the definitive study of pricing models under inflation, providing a solid basis for further research on this elusive question.
240 pp., 95 illus. $29.95 (May)

UNEMPLOYMENT AND MACROECONOMICS
Assar Lindbeck
Balancing theoretical insights with lessons drawn from the experience of many countries, Lindbeck examines employment and unemployment against the background of developed market economies during the past century.
The Ohlin Lectures 216 pp. $22.50 (June)

Original in Paperback
READINGS IN INTERNATIONAL BUSINESS
A Decision Approach
edited by Robert Z. Aliber and Reid W. Click
A solid grounding in the major practical business decisions that students and managers face in a global setting.
340 pp. $18.95 paper (April)

Original in Paperback
TAX POLICY AND THE ECONOMY
Volume 7
edited by James Poterba
Tax Policy and the Economy series 200 pp. $14.95 paper (May)

Now Available in Paperback
GEOGRAPHY AND TRADE
Paul Krugman
Eyskens Lecture Series 160 pp. $9.95 paper

REFORM IN EASTERN EUROPE
Olivier Blanchard, Rudiger Dornbusch, Paul Krugman, Richard Layard, and Lawrence Summers
128 pp. $8.95 paper

EXERCISES IN INTERTEMPORAL OPEN-ECONOMY MACROECONOMICS
Thomas H. Krueger and Jonathan D. Ostry
This exercise manual is a companion volume to *Fiscal Policies and the World Economy: An Intertemporal Approach*, Second Edition, by Jacob Frenkel and Assaf Razin.
150 pp. $15.00 paper (May)

To order call toll-free 1-800-356-0343 or (617) 625-8569 Fax orders (617) 625-6660 MasterCard and VISA accepted

The MIT Press
55 Hayward Street, Cambridge, MA 02142

NOW AVAILABLE IN PAPERBACK

INNOVATION AND GROWTH IN THE GLOBAL ECONOMY

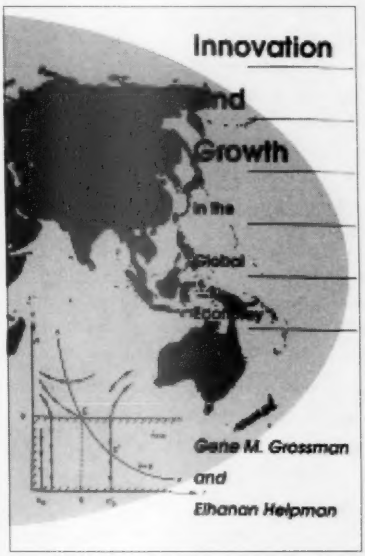
Gene M. Grossman and Elhanan Helpman

"A pathbreaking contribution by two of the smartest economists at the frontier of trade theory today." — Jagdish Bhagwati, Columbia University

Grossman and Helpman develop a unique approach in which innovation is viewed as a deliberate outgrowth of investments in industrial research by forward-looking, profit-seeking agents.

"More than two centuries after Adam Smith linked the extent of the market to the wealth of nations, Gene Grossman and Elhanan Helpman have developed the kind of coherent theoretical framework that previous discussions of trade, growth, development, and innovation have lacked. Any economist who wants to work on these timeless (and timely) issues should study this book."
— Paul M. Romer, University of California, Berkeley

359 pp. $16.95 paper



To order call toll-free 1-800-356-0343 or (617) 625-8569 MasterCard & VISA accepted

The MIT Press 55 Hayward Street, Cambridge, MA 02142

Economic Insight from Academic Press.

Statistical Methods in Econometrics

Ramu Ramanathan

Statistical Methods in Econometrics is appropriate for beginning graduate courses in mathematical statistics and econometrics in which the foundations of probability and statistical theory are developed for application to econometric methodology. Because econometrics generally requires the study of several unknown parameters, emphasis is placed on estimation and hypothesis testing involving several parameters. Accordingly, special attention is paid to the multivariate normal and the distribution of quadratic forms. Lagrange multiplier tests are discussed in considerable detail, along with the traditional likelihood ratio and Wald tests. Characteristic functions and their properties are fully explored. Also asymptotic distribution theory, usually given only cursory treatment, is discussed in detail.

January 1993, 405 pp., $55.00/ISBN: 0-12-576830-3

The Internationalization of the Firm

A Reader

edited by **Peter J. Buckley and Pervez N. Ghauri**

The Internationalization of the Firm is a collection of seminal articles on international business with particular reference to the experience of large European and American multi-nationals in Europe, Japan, and the developing countries. It covers the internationalization process in a chronological manner and analyzes the management of international operations. It will be a valuable text and resource for advanced under-graduates and MBA students.

Paperback: $29.95 (tentative)/ISBN: 0-12-139161-2
Casebound: $59.95 (tentative)/ISBN: 0-12-139160-4
January 1993, 400 pp.

Game Theory for Economists

Jürgen Eichberger

Game Theory for Economists introduces economists to the game-theoretic approach of modeling economic behavior and interaction, focusing on concepts and ideas from the vast field of game-theoretic models which find commonly used applications in economics.

Key Features

- Serves as a concise introduction to game theory which provides economists with the techniques and results necessary to follow the literature in economic theory
- Helps the reader formalize economic problems
- Concentrates on equilibrium concepts that are most commonly used in economics

July 1993, c. 330 pp./ISBN: 0-12-233620-8

Order from your local bookseller or directly from

ACADEMIC PRESS
HBJ Order Fulfillment Dept. **DM**17915
6277 Sea Harbor Drive, Orlando, FL 32887

CALL TOLL FREE
1-800-321-5068
FAX **1-800-336-7377**

Prices subject to change without notice. ©1993 by Academic Press, Inc. All Rights Reserved. KF/CEP/DC — 08033

"An outstanding tool for teaching econometrics."—Dr. Raouf S. Hanna
Dept. of Economics
Eastern Michigan University

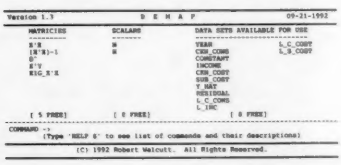
BEMAP™

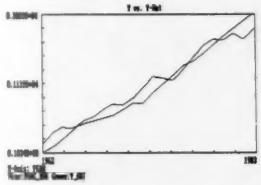
Basic Econometrics and Matrix Algebra Processor

An Easy-to-use, Matrix-oriented Mathematical Language for Econometric Applications and Instruction.

Featuring:

- Interactive and Batch Modes
- 16 Significant Digit Precision
- Full Color VGA Graphics (Up to five y-axis variables)
- Pre-written BEMAP programs for:
  - OLS Regression
  - AR1 Regression
  - Almon Lag (PDL)
  - Descriptive Statistics
  - Forecasting
  - Park's Test
- Comprehensive Instruction Manual
- "Introduction to Econometric Regressions using the BEMAP Language"





Requires IBM compatibility, 640K, VGA, and 3.5" Drive (5.25" on Request).

Order From: **Ann Arbor Econometrics**
Box 130661
Ann Arbor, MI 48113-0661

ONLY **$39.95** VERSION 1.3

For VISA, MasterCard, or P.O. Call:
1-800-952-8073

VISA MasterCard

Purchase Order by FAX

Does your econometric software package have all these features?

Linear Estimation: OLS, IV/2SLS, LIML, SUR, 3SLS, panel (fixed and random effects, unbalanced), complete regression diagnostics

Programming Features: Symbolic equation substitution and differentiation, matrix algebra equations, loops over numbers or names, procedures, conditional statements, storage of estimation results for further user computations

Data Input and Output: Databanks (TSP™ and microTSP™ format), formatted and unformatted READ and WRITE, spreadsheet format data input

Simulation: Model solution by Gauss-Siedel, Fletcher-Powell, or Newton's method, Box-Jenkins forecasting, Monte-Carlo simulation of normal, uniform, Poisson, or bootstrapped random variables

Nonlinear Estimation (using analytic gradients): FIML, TOBIT, PROBIT, LOGIT, GMM, minimum distance, sample selection models, general maximum likelihood

Time Series Estimation: ARIMA, GARCH-M, PDLs, Shiller lags, VAR, Kalman filter, AR1 and AR(p) regression

Available for almost any computer:

- IBM OS/MVS, IBM VM/CMS, VAX/VMS, SUN, RS/6000, DECstation, and most UNIX systems
- IBM PC/286/386/386SX/486 and compatibles (DOS-extended version, protected mode, for 286/386/386SX/486)
- Apple Macintosh II

Student versions, multiple copy discounts, and site licenses are also available.

On a 486/33Mhz computer, executing a panel data estimation (random and fixed effects model) with 25,000 observations and 10 independent variables takes only one minute. That's right: **one** minute.

If your software can't do all this, you need

TSP VERSION 4.2

For information, contact us at:

TSP International
P.O. Box 61015, Station A · Palo Alto, CA 94306 USA
Phone (415) 326-1927 · Fax (415) 328-4163

1993 ANNUAL MEMBERSHIP RATES

Membership includes a subscription to the *American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly). You may elect to receive two of the three journals; dues will be reduced by *$6.00 per year. Note: Regardless of your selection, you will receive the *Papers and Proceedings*.

JOURNAL OPTIONS AVAILABLE AT ANNUAL RENEWAL ONLY. Single copies may be purchased for $15.00 each after publication; prepayment required.

- Regular members with annual incomes of $30,000 or less $47.40
- Regular members with annual incomes above $30,000 but no more than $40,000 ..... $56.90
- Regular members with annual incomes above $40,000 ..... $66.40
- Junior members [available to registered students for five years maximum (eff. 1/1/90)]. Student status must be certified yearly by your major professor or school registrar $23.70
- **In Countries other than the U.S.A., Add $23.00 to cover postage.**
- Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association) $9.50

Please begin my membership in _____ **(year)** for the following period:

☐ **Jan.-Dec.** ☐ **April-March** ☐ **July-June** ☐ **Oct.-Sept.**

_____ _____ _____
First Name and Initial **Last Name** **Suffix**

_____
Address Line 1

_____
Address Line 2

_____
Address Line 3

_____
City

_____ _____
State or Country **Zip/Postal Code**

MAJOR FIELDS (TWO ONLY)
LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM *JEL*, "Classification System for Books."

☐ ☐

Phone Numbers: Hm _____ Ofc _____ Fax _____

Please type or print information above. Please pay with a check payable in United States Dollars. **Canadian and foreign payments must be by check drawn on a United States bank payable in United States Dollars.** Please note: It is the policy of the Association, not to refund membership payments.

Below for Junior Members Only

I certify that the person named above is enrolled as a student at _____

_____

_____
Authorized Signature

You will receive all three Association Publications unless indicated below.
I do NOT want to receive the following publication. Select only one.
(1) AER ☐
(2) JEL ☐
(3) JEP ☐

Dues (includes *AER, JEL* & *JEP.*) $_____

*Less $6.00 credit if applicable. (_____)
(I understand I will NOT receive the publication selected above.)

Optional donation. _____

Enter amount enclosed. $_____ USD

PLEASE SEND WITH PAYMENT TO:

American Economic Association
2014 Broadway, Suite 305
Nashville, Tennessee 37203
U.S.A.

World Economic Outlook Update, January 1993

The ***World Economic Outlook: Interim Assessment, January 1993,*** is a special update of the October 1992 edition, in light of the unexpected economic weakening in many key economies and a period of turmoil in the foreign exchange markets. It focuses mainly on the major industrial countries, but with significant economic implications for the rest of the world. The report examines the reasons behind the weakening and gives revised projections for growth and inflation and other economic variables, for industrial, developing, and former centrally planned economies.

ISBN 1-55775-278-8
Price: US$30.00 ($20.00 for university faculty and students).

Order from: Publication Services • Box No. XE-756
International Monetary Fund • 700 19th Street, N.W.
Washington, D.C. 20431, U.S.A. • Tel. (202) 623-7430
Fax (202) 623-7201

MAILING LIST CORRECTIONS

Information as it currently appears (attach mailing label if available). PLEASE PRINT.

FIRST NAME AND INITIAL | LAST NAME

ADDRESS LINE 1

ADDRESS LINE 2

ADDRESS LINE 3

CITY | STATE OR COUNTRY | ZIP/POSTAL CODE

TELEPHONE OFFICE | HOME | FAX

Please PRINT your NEW address here.

FIRST NAME AND INITIAL | LAST NAME

ADDRESS LINE 1

ADDRESS LINE 2

ADDRESS LINE 3

CITY | STATE OR COUNTRY | ZIP/POSTAL CODE

TELEPHONE OFFICE | HOME | FAX

Send to:

AMERICAN ECONOMIC ASSOCIATION
2014 Broadway, Suite 305 • NASHVILLE, TN 37203

BOYES/MELVIN
ECONOMICS

because these times demand change

▼ ▼ ▼ ▼ ▼ ▼ ▼ ▼ ▼

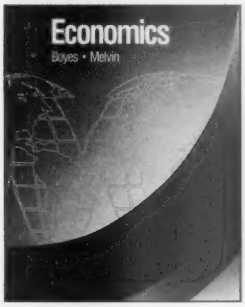
WILLIAM BOYES AND MICHAEL MELVIN
BOTH OF ARIZONA STATE UNIVERSITY

Complete Hardcover Edition

Two-Volume Paperback Edition:
Macroeconomics and
Microeconomics

Complete support package • 1991

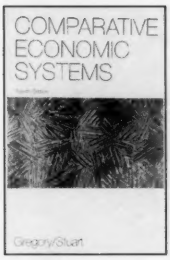
More instructors are switching to Boyes/ Melvin because it's the first introductory text created to present a global perspective right from the start. No other text integrates international coverage more thoroughly or provides



as many descriptive explanations, models, and examples. Organized to reflect today's economy, the macro section uses AD-AS as the operating model for applications and policy analysis, and the micro section features complete, separate chapters on the environment and on aging and health care. And teaching and learning couldn't be easier with outstanding features that demonstrate to students the critical role economics plays in their lives.

▼ ▼ ▼ ▼ ▼ ▼ ▼ ▼ ▼

ALSO AVAILABLE



Comparative Economic Systems, Fourth Edition
Paul Gregory, University of Houston
Robert C. Stuart, Rutgers University

The Fourth Edition gives students the tools they need to compare economic systems in today's rapidly changing environment.

528 pages • hardcover • Complete support package
1992

Business Statistics for Management and Economics, Sixth Edition
Wayne W. Daniel
James C. Terrell
Both of Georgia State University

Daniel and Terrell have enhanced the Sixth Edition with a lively, real-world emphasis and have integrated MINITAB commands and printouts throughout.

920 pages • hardcover • Complete support package
1992

For adoption consideration, request examination packages from your Houghton Mifflin regional office.

Houghton Mifflin

13400 Midway Rd., Dallas, TX 75244-5165 ▲ 1900 S. Batavia Ave., Geneva, IL 60134
925 E. Meadow Dr., Palo Alto, CA 94303 ▲ 101 Campus Dr., Princeton, NJ 08540

Dryden/HBJ
Excellence in Economics

New in 1993

Principles of Economics,
Second Edition
James F. Ragan, Jr.
Lloyd B. Thomas, Jr.

Money, Banking, and Financial Markets,
Second Edition
Meir Kohn

Macroeconomics: Analysis and Applications
Steven L. Green

Managerial Economics,
Seventh Edition
Mark Hirschey
James L. Pappas

Public Finance,
Fourth Edition
David N. Hyman

Labor and the Economy,
Third Edition
Howard Wachtel

History of the American Economy,
Third Edition
Robert C. Puth

The Road to Capitalism
David Kennett
Marc Lieberman

Also Available

Principles textbooks by:
William J. Baumol and Alan S. Blinder (Fifth Edition)
James D. Gwartney and Richard L. Stroup (Sixth Edition)
Edwin G. Dolan and David E. Lindsey (Sixth Edition)

Managerial Economics
William F. Samuelson
Stephen G. Marks

Intermediate Microeconomics
Michael B. Ormiston

Introductory Econometrics with Applications,
Second Edition
Ramu Ramanathan

Applied Econometrics: Problems with Data Sets
William F. Lott
Subhash C. Ray

Microeconomic Theory: Basic Principles and Extensions,
Fifth Edition
Walter Nicholson

Macroeconomic Decision Making in the World Economy: Text and Cases,
Third Edition
Michael G. Rukstad

Dryden
HBJ
The Dryden Press HBJ College Publishers
301 Commerce St. Ft. Worth, TX 76102

N. Gregory Mankiw

Macroeconomics

This thoroughly modern book for your intermediate course

- covers long-run classical theory before short-run Keynesian analysis
- offers a well balanced approach
- is clearly and concisely written, easy to read and understand
- is accurate and precise, without being overly technical
- has excellent case studies that shed light on real-world events
- can be covered in one term
- will delight both you and your students, as it does those who already use it at:

Alabama: University of Alabama, Birmingham **Alaska**: University of Alaska at Anchorage **Arizona**: University of Arizona **Arkansas**: University of Arkansas at Little Rock **California**: California State University at Bakersfield, at Fullerton, at Long Beach, and at Los Angeles, Chapman College, Claremont College, Harmelí College, Occidental College, Santa Clara University, Sonoma State University, Stanford University, University of California at Berkeley, at Davis, at Irvine, at Los Angeles, at Riverside, and at San Diego-La Jolla, University of San Diego, University of Southern California, University of the Pacific **Colorado**: Colorado State University, Metropolitan State College, Regis University, University of Colorado at Boulder **Connecticut**: Fairfield University, University of Connecticut at Storrs, Yale University **Delaware**: University of Delaware **District of Columbia**: American University, Catholic University, Georgetown University **Florida**: Edison Community College, Jacksonville University, Rollins College, University of Central Florida, University of South Florida, University of Tampa **Georgia**: Berry College, Emory University, Georgia State University **Hawaii**: University of Hawaii at Honolulu **Idaho**: Boise State University **Illinois**: Augustana College, DePaul University, Illinois Central University, Illinois State University, Northern Illinois University, Northwestern University, Principia College, Sangamon State University, Southern Illinois University at Carbondale and at Edwardsville, University of Illinois at Chicago and at Urbana-Champaign **Indiana**: Indiana-Purdue University at Indianapolis, Indiana University at Bloomington, University of Evansville, University of Notre Dame **Iowa**: Iowa State University, Simpson College, University of Iowa, University of Northern Iowa **Kansas**: Fort Hays State University, Kansas State University, University of Kansas, Wichita State University **Kentucky**: University of Kentucky **Louisiana**: Louisiana State University at Alexandria and at Baton Rouge, Tulane University, University of New Orleans **Maine**: Bates College, Bowdoin College, Colby College **Maryland**: Towson State University, University of Maryland at College Park **Massachusetts**: Amherst College, Bentley College, Boston University, Bridgewater State University, Clark University, Fitchburg State College, Gordon College, Harvard University, Mount Holyoke College, Salem State College, Smith College, Stonehill College, Tufts University, University of Massachusetts at Dartmouth, Wellesley College, Williams College **Michigan**: Albion College, Michigan State University, Northern Michigan University, University of Detroit, University of Michigan at Ann Arbor **Minnesota**: Augsburg College, Mankato State University, University of Minnesota at Twin Cities, Winona State University **Mississippi**: Alcorn State University **Missouri**: Northwest Missouri State University, St. Louis University, Southeast Missouri State University, University of Missouri at St. Louis, Washington University **Montana**: Montana State University, University of Montana **Nebraska**: Creighton University, University of Nebraska at Lincoln **Nevada**: University of Nevada at Las Vegas **New Jersey**: Rutgers University at Newark and at New Brunswick, Seton Hall University, Stevens Institute of Technology **New Mexico**: New Mexico State University, University of New Mexico at Albuquerque **New York**: Binghamton University, Brooklyn College, Cornell University, Ithaca College, Manhattan College, New York University, Nyack College, Rensselaer Polytechnic Institute, Skidmore College, State University of New York at Albany, at Buffalo, at Oswego, at Potsdam, and at Stony Brook, St. John's University, Syracuse University, Union College, University of Rochester **North Carolina**: Duke University, East Carolina University, University of North Carolina at Chapel Hill and at Charlotte **North Dakota**: University of North Dakota at Bismarck and at Grand Forks **Ohio**: Bowling Green State University, Cleveland State University, Denison University, Ohio State University at Columbus, Ohio University at Athens and at Lancaster, University of Cincinnati, University of Toledo, Wright State University, Youngstown State University **Oklahoma**: University of Central Oklahoma, University of Oklahoma, University of Tulsa **Oregon**: Lewis and Clark College, Western Oregon State College **Pennsylvania**: Bucknell University, Carnegie-Mellon University, Drexel University, Indiana University of Pennsylvania, Lafayette College, Lehigh University, Pennsylvania State University at Delaware County, at Ogontz, and at University Park, Philadelphia College of Textiles, Temple University, University of Pittsburgh at Johnstown and at Pittsburgh, Villanova University, Westminster College **Rhode Island**: Brown University, Bryant College, University of Rhode Island **South Carolina**: College of Charleston, University of South Carolina at Columbia **South Dakota**: Black Hills State College **Tennessee**: Fisk University, Memphis State University, University of Tennessee at Knoxville, Vanderbilt University **Texas**: Angelo State University, Baylor University, Lamar University, Southern Methodist University, Southwest Texas State University, Southwestern University, Texas A & M University, Texas Christian University, Texas Technological University, University of Houston at University Park, University of St. Thomas, University of Texas at Arlington, at Austin, and at Dallas, West Texas State University, Utah, Brigham Young University **Vermont**: Middlebury College, St. Michael's College **Virginia**: College of William and Mary, Foreign Service Institute, George Mason University, James Madison University, Old Dominion University, University of Virginia, Virginia Polytechnic Institute and State University **Washington**: Eastern Washington University, Pacific Lutheran University, University of Washington **West Virginia**: Davis and Elkins College **Wisconsin**: Carthage College, Marquette University, Ripon College, University of Wisconsin at Eau Claire, at Green Bay, at Madison, at Milwaukee, and at Platteville **Wyoming**: Eastern Wyoming College **Canada**: Carleton University, Guelph University, Laurentian University, Queens University, University of Calgary, University of Montreal, University of New Brunswick, University of Toronto, University of Western Ontario, Wilfrid Laurier University, York University **Australia**: La Trobe University, Macquarie University, Queensland University of Technology, University College of Central Queensland, University of South Australia **Belgium**: Universite Libre de Bruxelles **Cyprus**: Cyprus College **Denmark**: Roskilde University Center, University of Copenhagen **Finland**: Helsinki School of Economics, University of Helsinki, University of Joensuu, University of Turku **France**: Institut D'Etudes Politiques, University of Caen, University of Pans I **Germany**: University of Augsburg, University of Hamuburg, University of Osnabrucle, University of Saarbrucken **Hong Kong**: Chinese University of Hong Kong **Iceland**: University of Iceland **Ireland**: University College at Cork, University College at Dublin, University of Dublin – Trinity College **Italy**: Universita L'Bocconi **Korea**: Korea University **Netherlands**: Tilburg University, University of Amsterdam, University of Groningen **Norway**: Norwegian School of Management, University of Bergen, University of Oslo **Philippines**: Center for Research and Communications **South Africa**: University of Durban-Westville **Sweden**: Ostersund University, Stockholm University, University of Umek **Switzerland**: University of Basel, University of St. Gallen **United Kingdom**: Dundee Institute of Technology, Kingston University, London School of Economics, Newcastle Polytechnic, Sheffield University, University of Birmingham University of Essex, University of Hull, University of Reading, University of St. Andrews, University of Southampton, University of Wales at Bangor

Worth Publishers

33 Irving Place, New York, New York 10003

212-475-6000 or 800-223-1715

The Macroeconomic Picture of the '90s

Trade and Development Report, 1992

"The world economy has been suffering its most severe recession since the Second World War. No single country can solve the macroeconomic problem on its own: the situation demands improved coordination."

This annual report provides a thorough review of the world economic situation and reviews the significant trends and developments in international trade. It examines in:-

Part I -

The Current situation and Prospects - analysing the recent developments and short-term prospects of international markets; regional performance and short-term prospects of developing countries, China, Central and Eastern Europe and the former USSR; developed market-economy countries; and gives a long-term perspective on development and growth.

Part II -

Finance and Economic Expansion - examining recent trends, financial positions, renegotiation and reduction of bank debt, costs and other terms of private export credits and official debt reschedulings of the international capital markets and external debt; and the undercurrent of debt deflation, showing current conditions in financial markets of the United States and other countries, sources of current difficulties and the likely consequences for recovery.

Part III -

Policy Reform in Developing Countries - examining the issues and the evidence from the 1970s and 1980s, inward and outward orientation in development and the phasing of reform; and discusses the characteristics, performance and factors affecting performance in reforming public enterprises. It offers insight into areas of reform and draws some conclusions.

E.92.II.D.7 92-1-112321-6 $42.00

Industry and Development, 1992/93

"The importance of privatization as a favoured policy tool cannot be denied."

This year's Global Report has thus taken up 'Privatization' as its special theme in an endeavour to help policy makers contemplate privatization for the first time as well as to reassure those who have embarked on that course and might be encountering problems. A chapter is dedicated to reviewing the recent privatization experiments in some 50 countries. The analysis is given an added dimension by answers from a questionnaire survey of managers of industrial enterprises facing the challenge of privatization.

It also offers a wealth of information in its analysis of:

- *the industrial performance of major economies around the world*
- *eight Asian economies which demonstrate the symbiotic relationships between manufacturing and agriculture and between manufacturing and services*
- *twelve selected industrial sectors based on contributions from editors of professional trade journals*

Researchers and policy makers will find the statistical annex a valuable source of information.

E.92.III.E.4 92-1-106276-4 $55.00 548pp

Send orders to:

United Nations Publications, Sales Section,
Room DC2-0853, **Dept. 016A,**
New York, N.Y. 10017, Tel. (800) 253-9646,
(212) 963-8302, Fax. (212) 963-3489.

Visa, MasterCard and American Express accepted for orders over $15.00.
Please add 5% of gross ($3.50 minimum) for shipping & handling.

United Nations Publications

HEALTH CARE AND PENSION STUDIES FROM THE OECD

U.S. Health Care at the Cross-Roads.

(Health Policy Studies No. 1).

U.S. Health Care costs are almost twice the average of other industrialized nations on a per capital basis. As health care moves to the top of the policy agenda in the United States, two issues—high and rising costs and inadequate access to care—have defined the policy debate. This report, drawn from the OECD Economic Survey of the United States 1992, analyzes health-care problems in the United States from an international perspective. The report concludes that real policy options for containing cost and increasing access are much narrower than they appear, and should be of interest to health care professionals, insurers, and anyone concerned with the direction of health care in the United States.

11-92-03-2, November 1992, 99 pages, ISBN 92-64-13780-7, $15.00

Private Pensions and Public Policy.

Governments have begun to look to the private sector for ways in which retirement income might be provided without raising taxes in the years ahead, and which might also reduce pressures on the public budget in the short term. Often lacking, however, is a comprehensive view of all the roles played by private pensions, how well those roles are performed, what conflicts may be engendered between public and private goals, what potential private pensions may hold—and at what cost. This volume is a contribution to informed debate on the issue.

81-92-06-1, December 1992, 160 pages, ISBN 92-64-13790-4, $35.00

Private Pensions in OECD Countries: the United States.

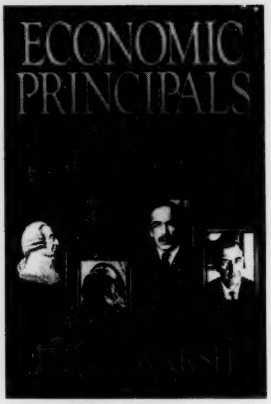
Interest in changing the balance between public and private provision of retirement income, with increased reliance on private sources, is emerging in OECD countries. Although documentation and analysis of public systems are abundant, much less is known about the performance of private plans. This volume, one of a series of national case studies, describes the characteristics of private arrangements in the United States. It discusses their prevalence, structure, financing and administration, tax treatment, regulation, and their effects on the distribution of retirement income.

81-93-01-1, February 93, 100 pages, ISBN 92-64-13802-1, $22.00

To order, send your check or money order to:

OECD Publications and Information Center
2001 L Street, NW, Washington, DC 20036
Telephone: (202) 785-6323
Book Orders: 800-456-OECD

NEW FROM THE FREE PRESS



ECONOMIC PRINCIPALS

The Masters and Mavericks of Modern Economics

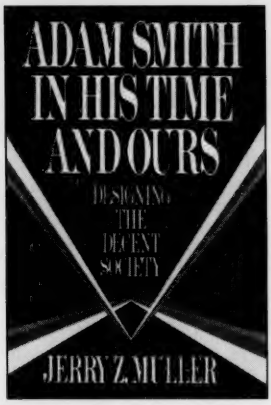
David Warsh

"David Warsh's column in the *Boston Globe* is unique: the only regular, thoughtful, and frequently irreverent discussion of economics and economists in any American newspaper. A large number of readers will be delighted to have available this wonderful and most timely collection of essays."
—**Henry Rosovsky,** Harvard University

"Over the years no journalist has kept up so well and so accurately with the development of economic thought as David Warsh. His precise and interesting sketches have stood the test of years very well."
—**Kenneth J. Arrow,** Stanford University

"David Warsh has two qualities rare in economic journalism: he actually likes economics, and he knows a great deal about it. The columns contained in *Economic Principals* are consistently well-informed, lively, and interesting. Together, they provide a marvelous introduction to the current scene in economic thinking."
—**Robert E. Lucas, Jr.,** University of Chicago

1993 0-02-933996-0 $22.95



ADAM SMITH IN HIS TIME AND OURS

Designing the Decent Society

Jerry Z. Muller, *Catholic University*

"Brilliant and compelling... Particularly illuminating is its pervasive theme on the importance of unanticipated consequences in the workings and development of society and economy. By enlarging upon its text and subtext, Muller has placed *The Wealth of Nations* in its own time while enlarging upon its civil relevance for our time."
—**Robert K. Merton,** Columbia University

1992 0-02-922234-6 $22.95

THE SEVEN FAT YEARS

And How to Do It Again

Robert L. Bartley, Editor, *The Wall Street Journal*

"Bartley takes 'supply side' theory from the realm of mystical gospel for a few to a thoughtful exposition for the rest of us. In doing so, he has not only written a book that deserves to be read, but one that will influence thinking about public policy."
—**Paul Volcker**

1992 0-02-901915-X $22.95

THE CATHOLIC ETHIC AND THE SPIRIT OF CAPITALISM

Michael Novak,
American Enterprise Institute

"An original and exciting exploration of the moral structure of the free market economy, an excellent introduction to the catholic ethic (with a small 'c'), and a social and economic manifesto for a new approach to overcoming problems of poverty, ethnicity, and race relations."
—Jack Kemp

1993 0-02-923235-X $22.95

GENERATIONAL ACCOUNTING

Knowing Who Pays—and When—for What We Spend

Laurence J. Kotlikoff, *Boston University*

"Kotlikoff has something important to say. He makes an urgent case for examining just who pays for what we spend."
—The Wall Street Journal

1992 0-02-918463-0 $22.95

THE WINNER'S CURSE

Paradoxes and Anomalies of Economic Life

Richard H. Thaler, *Cornell University*

"Every now and then an economist writes a book that is both interesting and entertaining. Thaler has done just that."
—The Wall Street Journal

1992 0-02-932465-3 $24.95

THE COMPETITIVE ADVANTAGE OF NATIONS

Michael E. Porter,
Harvard Business School

"An all-embracing view of economic change that amounts in the end to a powerful analytic framework…A real achievement."
—The Economist

1990 0-02-925361-6 $39.95

Now in paperback—

CAPITAL IDEAS

The Improbable Origins of Modern Wall Street

Peter L. Bernstein

"An engaging history of the rise of investment theory, from French mathematician Louis Bachelier's dissertation, 'The Theory of Speculation,' to the formulation of portfolio insurance."
—USA Today

1993 0-02-903012-9 $12.95 paper

Named two of the five most cited books in the field of economics—

THE ECONOMIC INSTITUTIONS OF CAPITALISM

Firms, Markets, Relational Contracting

Oliver E. Williamson,
Yale University

1987 0-02-934821-8 $17.95 paper

MARKETS AND HIERARCHIES

Analysis and Antitrust Implications

Oliver E. Williamson,
Yale University

1983 0-02-934780-7 $14.95 paper

For Visa, MasterCard, Discover, or American Express orders, call toll-free 1-800-323-7445
Monday through Friday between 9am-9pm Eastern time.

THE FREE PRESS

A Division of Macmillan, Inc./A Maxwell Macmillan Company
866 Third Avenue, New York, NY 10022

Announcing A New Journal.....

ECONOMIC DESIGN

Editor-in-Chief:
Murat Sertel, *Bogaziçi University, Turkey*
Co-Editor:
William Thomson, *University of Rochester, USA*
Honorary Editor:
Leonid Hurwicz, *University of Minnesota, USA*

Associate Editors

A Board of approximately 25 Associate Editors will be appointed. At this moment the Board consists of:
Jeffrey Banks, *University of Rochester, USA*
Salvador Barbera, *Universidad Autónoma de Barcelona, Spain*
Bhaskar Dutta, *Indian Statistical Institute, India*
Faruk Gül, *Stanford University, USA*
Matthew Jackson, *Northwestern University, USA*
Richard Kihlstrom, *University of Pennsylvania, USA*
Paul R. Kleindorfer, *University of Pennsylvania, USA*
Semih Koray, *Middle East Technical University, Turkey*
Eric S. Maskin, *Harvard University, USA*
Thomas R. Palfrey, *California Institute of Technology, USA*
Stanley Reiter, *Northwestern University, USA*
John E. Roemer, *University of California, Davis, USA*
Ariel Rubinstein, *University of Tel Aviv, Israël*
David Schmeidler, *University of Tel Aviv, Israël*
Sanjay Srivastava, *Carnegie Mellon University, USA*
Joseph E. Stiglitz, *Stanford University, USA*
Kotaro Suzumura, *Hitotsubashi University, Japan*

Aims and Scope

In this Age of Economic Design, the accumulated traditions and wealth of knowledge in normative and positive economics and the strategic analysis of Game Theory are applied with novel ideas in the creative tasks of designing and assembling diverse legal-economic instruments. These include constitutions and other assignments of rights, mechanisms for allocation or regulation, tax and incentive schemes, contract forms, voting and other choice aggregation procedures, markets, auctions, organizational forms, such as partnerships, together with supporting membership and other property rights, and information systems. These designs, the methods of analysis used in their scrutiny, as well as the mathematical techniques and empirical knowledge they employ, along with comparative assessments of the performance of known economic systems and implemented designs, all of these form natural components of the subject matter of **Economic Design.**

Call for Papers

If you feel that your manuscript will fit the scope of the journal **Economic Design**, please send four copies to the editor:
Professor M. Sertel
Department of Economics
Bogaziçi University
Bebek, Istanbul, Turkey.

Publication Schedule

The first issue of **Economic Design** is scheduled for publication at the beginning of 1994. One volume of 4 issues will be published in 1994.

NORTH-HOLLAND
(An Imprint of Elsevier Science Publishers)

ELSEVIER SCIENCE PUBLISHERS
Attn: Heleen A.C. Gierveld
P.O. Box 1991
1000 BZ Amsterdam, The Netherlands

in the USA/Canada:
Attn: Judy Weislogel
P.O. Box 945, Madison Square Station
New York, NY 10160-0757

419/NECONDES1

INTERNATIONAL ECONOMIC JOURNAL

VOLUME 6 WINTER 1992 NUMBER 4

The Role of International Trade in the Convergence of Per Capita GDP in the OECD: 1950-85 — Farhad Rassekh
Effect of Risk Aversion on the Incentive to Share Information — H. S. Hwang and N. S. Lee
International Migration and Welfare in the Presence of Non-Traded Goods: An Interconnected Markets Approach — Bharati Basu
Intra-Industry Trade: The Australian Experience — R. Ratnayake and P. Athukorala
Factor Price Shocks, Factor Substitution and Their Implication for Policy — Z. Mahmood
Protection, Terms of Technology Transfer and Foreign Investment: A Welfare Analysis — M. Anam and A. B. Supapol
The Cost of Capital and the Multi-Nationalization of Japanese Banking — John Thornton
Changes in the Extent of Financial Interdependence between the G7 Countries in the 1970s and 1980s — Mark J. Holmes and Eric J. Pentecost

MANUSCRIPTS SUBMITTED (*in triplicate*) *in Asia for publication should be sent to the Managing Editor, Professor Wontack Hong, Department of International Economics, College of Social Sciences, Seoul University, Seoul 151-742, Korea. Manuscripts from countries outside Asia should be sent to the Co-Editor, Professor Young Chin Kim, Department of Economics, Northern Illinois University, DeKalb, Illinois 60115, U.S.A. Personal subscription price of a volume (which includes postage) is $40 per year. Institutional subscription price is $50 per year. Cheques should be made payable to International Economic Journal and sent directly to Managing Editor. (Volumes 1~5 are available.)*

THE REVIEW OF FUTURES MARKETS — NOW AVAILABLE

Volume 10, Number 2 contains seven research studies on topics of current importance.

The Review of Futures Markets is a fully refereed, professional journal on futures and options. The cost of a subscription is $35 for all 3 issues, or $15 for a single issue. To order, complete the form below. Include your Visa or MasterCard number, or a check or money order made payable to the Chicago Board of Trade, and mail to: Chicago Board of Trade, Publications Department, 141 West Jackson Blvd., Suite 2210, Chicago, IL 60604-2994

Visa/MasterCard orders can also be made by phone; call 1-800-THE-CBOT, ext. 31 or 312-435-3558, ext. 31.

ORDER FORM (check one)

U.S. only	Total
☐ Volume 10 (3 issues)	$35.00
☐ Volume 10, Number 2	$15.00
Orders shipping outside the U.S. (Payment should be drawn on a U.S. bank ($))	
☐ Volume 10 (Airmail)	$55.00
☐ Volume 10 (Surface)	$42.50
☐ Volume 10, #2 (Airmail)	$27.00
☐ Volume 10, #2 (Surface)	$20.00

Visa/MC#_____

Exp. Date:_____/_____ (Month/Year)

Cardholder Name:_____

Ship To: (Please type or print clearly)

Name

Company

Address

Address

City, State, Zip

Country

( ) Telephone Number

AER 31

NEW FROM

GPG GREENWOOD PUBLISHING GROUP INC.

Economic Activity, Trade, and Industry in the U.S.—Japan-World Economy
A Macro Model Study of Economic Interactions
By F. Gerard Adams, Byron Gangnes, and Shuntaro Shishido

A pioneering effort to link quantitatively the economies of the U.S. and Japan through a large-scale econometric model.

Praeger Publishers. 1993. 0-275-94488-3. $46.00E.

The Age of Giant Corporations
A Microeconomic History of American Business, 1914-1992, A Third Edition
By Robert Sobel

A popular and valuable text, The Age of Giant Corporations is now available in a third edition, bringing the history up to 1992 and including a new chapter on the era of junk bonds and its aftermath.

Greenwood Press. 1993. 328 pages.
0-313-28730-9. $49.95.
A paperback is also available:
0-275-94470-0. $18.95

Corporate Environmentalism in a Global Economy
Societal Values in International Technology Transfer
By Halina Szejnwald Brown, Patrick Derr, Ortwin Renn, and Allen L. White with a chapter by Jeanne X. Kasperson and Roger E. Kasperson

Analyzes values and perceptions relevant to multinational corporations' hazardous facility siting in developing countries.

Quorum Books. 1993. 264 pages.
0-89930-802-3. $49.95.

ANZUS Economics
Economic Trends and Relations among Australia, New Zealand, and the United States
Edited by Richard W. Baker and Gary R. Hawke

Examines the changing economic nature of the ANZUS Alliance.

Praeger Publishers. 1992. 276 pages.
0-275-94381-X. $49.95.

Multinational Culture
Social Impacts of a Global Economy
Edited by Cheryl R. Lehman and Russell M. Moore

The result of a conference on multinational culture, this collection explores the sociocultural impacts of a global economy. Examined are: government policies & their relationship to multinational activities; the formation & regulation of international capital; labor market segmentation & protectionism; managing multinationals without sacrificing ethical standards or profits; environmental impacts; & other aspects of a global economy.

Greenwood Press. 1992. 360 pages.
0-313-27822-9. $55.00.

America and the Multinational Corporation
The History of a Troubled Partnership
By John J. Reardon

Tells the story of the postwar relationship between America's multinational corporations and the U.S. government.

Praeger Publishers. 1992. 194 pages.
0-275-93918-9. $45.00.

ORDER TOLL-FREE, 24 HOURS-A-DAY: 1-800-225-5800
Please call us for a free catalog at (203) 226-3571

Q

GREENWOOD PRESS • PRAEGER PUBLISHERS • QUORUM BOOKS

88 Post Road West, Box 5007, Westport, CT 06881-5007
AD 93 (203) 226-3571 • FAX (203) 222-1502

Journal of Forensic Economics

A NEW JOURNAL IN A NEW FIELD

A tri-annual, refereed academic journal which publishes both theoretical and applied papers concerning research and techniques in forensic economics. Articles encompass topics of interest to economists, attorneys and vocational experts engaged in such fields as personal injury/wrongful death litigation, business valuation, employment discrimination, commercial litigation, and anti-trust litigation. Once each year bibliographies and a listing of members are published.

The JFE is published by the NATIONAL ASSOCIATION OF FORENSIC ECONOMICS, a non-profit association of economists, financial analysts, rehabilitation specialists, and attorneys in all 50 states as well as Canada, England, Germany, Hong Cong, Ireland, Japan, Puerto Rico, South Africa and the Virgin Islands. NAFE is not an accrediting association, but is a professional organization designed to foster scholarly debate in issues of forensic economics. NAFE membership includes a subscription to the JFE, participation in regional and national NAFE programs, a membership directory, quarterly newsletters informing our members of new research in forensic economics, new applications of economics in litigation, new papers being presented, upcoming meetings and calls for papers.

MEMBERSHIP AND ADDITIONAL INFORMATION:

National Association of Forensic Economics
Post Office Box 30067
Kansas City, MO 64112
(816) 235-2833 FAX: (816) 235-5263

AP

Advanced Placement
Dedicated to Educational Excellence
for Nearly Four Decades

Faculty Consultants for the AP Reading

Each year, the College Board's Advanced Placement (AP™) Program gives hundreds of thousands of exceptionally able high school students an opportunity to take rigorous college-level courses and appropriate exams in 16 disciplines. More than 2,900 colleges and universities offer credit or advanced standing to students based on their exam performance. For six days in June, more than 2,400 college faculty and AP teachers from across North America gather at Trenton State College, Clemson University and Trinity University at the annual AP Reading to evaluate and score students' essays. The participants also exchange ideas and contribute suggestions about their discipline, their courses, and the AP Examinations. Participants are paid honoraria, provided with housing and meals, and reimbursed for travel expenses.

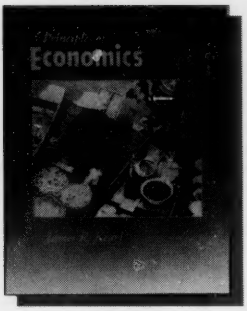
Applications are now being accepted for Faculty Consultants to the College Board's Advanced Placement Readings in Economics. Applicants should currently be teaching or directing instruction for the first-year college course in either Microeconomics or Macroeconomics.

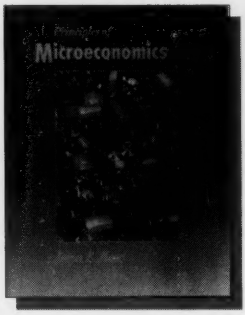
For an application or additional information, please contact: Dr. Walt MacDonald, Director, Advanced Placement Program, Educational Testing Service, Rosedale Rd. 85-D, Dept. 93-19, Princeton, NJ, 08541. ***Educational Testing Service is an Equal Opportunity/Affirmative Action Employer and especially encourages minorities and women to apply.***

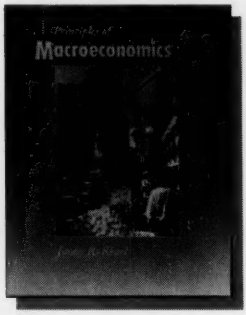
Activate Your Students' Interest!

Principles of Economics

James R. Kearl, Brigham Young University







These texts are accompanied by unique problem-solving workbooks by Mark Rush! Not the standard books of fill-in-the-blanks, these singular workbooks offer extensive problem sets developed directly from recent data and articles in *The Wall Street Journal* and the *New York Times*. Kearl and Rush's "learning-by-doing" approach treats economics as a tool to be used, not just memorized.

Features of all three texts

- Each chapter ends with fully solved sample problems (both qualitative and quantitative).
- Incentives and behavior is treated explicitly as a core tool in Chapters 1 and 2, then developed throughout.
- Singular international coverage begins with trade and exchange rate theory in Chapter 4.

Features of *Microeconomics*

- No text better explains how markets clear, with a stronger emphasis on supply and demand, relative pricing elasticity, and substitution effects.
- Law and economics is stressed throughout; for example, tort law appears as an alternative to economic regulation.

Features of *Macroeconomics*

- This text is the first to devote the beginning macro chapter almost exclusively to growth and instability; growth is then integrated throughout the text.
- Extensive applications cover the current recession/stagnation.
- By presenting exchange rates up front, Kearl is better able to explain the international macro events reflected daily in the news, such as the effects of German–U.S. interest rate differentials.

For more information, call your D. C. Heath representative toll-free: **1-800-235-3565**.

D C Heath

A Raytheon Company

College Division • 125 Spring Street • Lexington, MA 02173

Playing by the Rules

Games and Economic Behavior

Editor
Ehud Kalai
Northwestern University
Evanston, Illinois

Games and Economic Behavior . . .

- Publishes original and survey papers dealing with game-theoretic modeling in the social, biological, and mathematical sciences
- Presents papers that are mathematically rigorous as well as accessible to readers in related fields
- Facilitates cross-fertilization between the theory and application of game-theoretic reasoning

Volume 5 (1993), 4 issues
In the U.S.A. and Canada: $133.00
All other countries: $161.00
ISSN 0899-8256

New

Journal of Housing Economics

Editor
Henry O. Pollakowski
Joint Center for Housing Studies
Harvard University, Cambridge
Massachusetts

Journal of Housing Economics . . .

- Provides a focal point for the publication of economic research related to housing
- Encourages papers that bring to bear careful analytical technique on important housing-related questions
- Covers the broad spectrum of topics and approaches which comprise housing economics, including analysis of important public policy issues

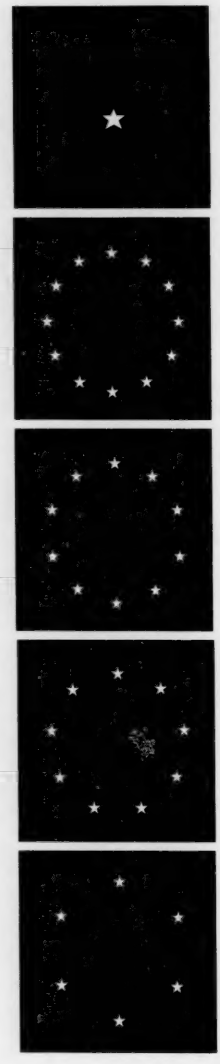
Volume 3 (1993), 4 issues
In the U.S.A. and Canada: $110.00
All other countries: $135.00
ISSN 1051-1377

Sample copies and privileged personal rates are available upon request.
For more information, please write or call:

ACADEMIC PRESS, INC., Journal Promotion Department
1250 Sixth Avenue, San Diego, CA 92101, U.S.A.
(619) 699-6742

All prices are in U.S. dollars and are subject to change without notice.
Canadian customers: Please add 7% Goods and Services Tax to your order.

S3228



Singular Europe

ECONOMY AND POLITY OF THE EUROPEAN COMMUNITY AFTER 1992

William James Adams, Editor

CONTRIBUTORS:

Severin Borenstein • Air Travel
Bo Carlsson • Small Business
Gunter Dufey • Banking
Reinhard Ellger • Telecommunications
John H. Jackson • World Trade
Alexis Jacquemin • Corporate Strategy
Frédéric Jenny • Competition
Stephan Leibfried • Social Policy
Ernst-J. Mestmäcker • Germany
Tommaso Padoa-Schioppa • Monetary Union
Gary R. Saxonhouse • Japan
John D. Steinbruner • United States
William Wallace • Foreign Policy
Joseph H. H. Weiler • Political Union
Susan L. Woodward • United States

The future impact of Program 1992 upon the European and world economies is discussed by leading international experts.
cloth $45.00 / £30.00

Sponsored by the Chicago Council on Foreign Relations

The University of Michigan Press
839 Greene Street, PO Box 1104
Ann Arbor, Michigan 48106-1104
Plymbridge Ltd.
Estover, Plymouth PL6 7JA UK

Journal of Economic Integration

Solicits Papers to Compete for the Annual Daeyang Prize

Volume 7, No. 1 Spring 1992

Ronald W. Jones, *Factor Scarcity, Factor Abundance and Attitudes Towards Protection: The 3×3 Model*

Edward Tower, *Domestic Resource Cost*

Govind Hariharan and Howard J. Wall, *Intertemporal Optimization under Threat of VER*

Robert G. Murphy, *The Terms of Trade, Investment, and the Current Account*

Marcelo Bianconi, *Monetary Growth Volatility and Asset Prices in a Two-Country Cash-in-Advance Model*

Myung-gun Choo, *Provisional PARFTA (Pacific Rim Free Trade Agreement) and Its Economic Effect*

The Journal is published biannually (in Spring and Autumn) by the Institute for International Economics, Sejong University, Seoul, Korea.

Journal of Economic Integration will publish articles that investigate economic issues in international trade and finance. While the emphasis is on the broad issue of international economic integration, the Journal is interested in scholarly articles on applied and theoretical issues of international economics. Thus, the Journal intends to provide a forum for theorists, empirical researchers, and policy makers.

- From the papers selected for publication, the Prize committee will choose the best manuscript to receive the US$5,000 Daeyang Prize. The winner of the Prize is announced in the Spring issue each year.
- Manuscripts should be in English, typed in double spacing with wide margins (3 cm) on one side of standard paper. Footnotes, figures and tables should be on separate sheets.
- Manuscripts should be sent in triplicate to the Managing Editor, Institute for International Economics, Sejong University, Seoul, Korea.
- Subscription prices are US$20 per year for individuals and US$30 per year for institutions. All issues are normally sent by surface. Additional US$5 per year is required for air delivery. Checks and money orders are payable to the Sejong University.

Institute for International Economics
Sejong University
Seongdong-ku, Seoul, Korea 133-747 Tel: (02) 460-0338 Fax: (02) 460-0200

How Many Candles Were On Your Cake The Last Time You Thought About Buying Insurance?

Face it — it's been a long time. Styles have changed. So has your family, maybe even your job. And most likely, the insurance you bought then isn't enough to cover your family today. That's why you need insurance that you can easily update as your life changes — AEA Term Life Insurance.

Designed For You.

Finding insurance that's right for you can be difficult and time consuming. But Seabury & Smith has done that work for you. Through their research, they've found a Term Life Plan designed to help members of our profession provide for the future of their families. What's more, you can be sure the plan is constantly being evaluated to better meet the needs of AEA members.

Flexible.

Updating your insurance doesn't have to be a hassle. With our Term Life Plan, as your needs change, so can your coverage. Insurance through your association is designed to grow with you—it even moves with you when you change jobs.

Affordable.

Seabury & Smith offers AEA members the additional benefit of reasonable rates, negotiated using our group purchasing power. Find out more about AEA Term Life insurance; call 1 800 424-9883 (in Washington, DC, (202) 457-6820) between 8:30 a.m. and 5:30 p.m. Eastern Time.

AEA Insurance

Designed for the way you live today.
And tomorrow.
Administered by Seabury & Smith.

MicroTSP... the leader in econometric software

MicroTSP 7.0 for PC Compatibles

The best selling econometric software is now even better. New features include a wide range of model evaluation & hypothesis tests, recursive estimation, page layout features in graphics, forecast confidence intervals, and much more.

MicroTSP 6.6 for the Macintosh

This all new version was designed to make full use of the Apple Macintosh. View and manipulate multiple windows of statistical output and graphics using the standard Macintosh interface. Data capacity is limited only by available memory.

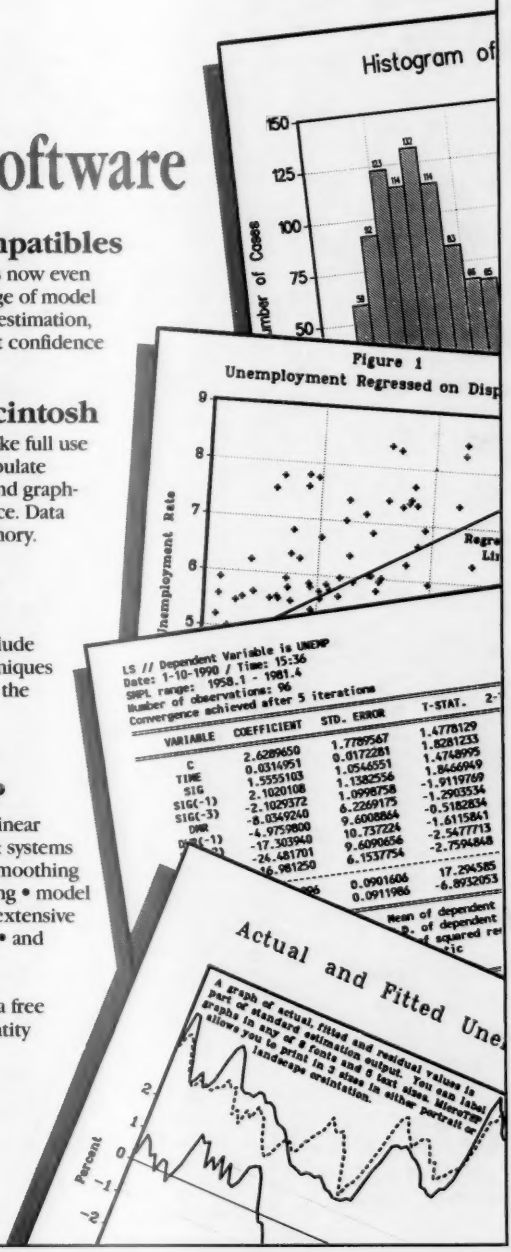
Student Versions for PC and Macintosh

These limited data capacity versions include all of the econometric and graphic techniques of MicroTSP 6.5. Price is comparable to the average textbook.

Standard Features of All Versions of MicroTSP

Menu, command, or batch operation • linear and nonlinear estimation of equations & systems • Box-Jenkins • ARMAX • exponential smoothing • logit • probit • PDLs • VAR • forecasting • model simulation • handling of missing data • extensive high resolution graphics • data transfer • and much more.

Write or call for more information and a free demonstration disk. Academic and quantity discounts are available.


Histogram of
Number of Cases
Figure 1
Unemployment Regressed on Disp
Unemployment Rate
LS // Dependent Variable is UNEMP
Date: 1-10-1990 / Time: 15:36
SMPL range: 1958.1 - 1981.4
Number of observations: 96
Convergence achieved after 5 iterations
VARIABLE COEFFICIENT STD. ERROR T-STAT.
C 2.6289650 1.7789567 1.4778129
TIME 0.0314951 0.0172281 1.8281233
SIG 1.5555103 1.0546551 1.4748995
SIG(-1) 2.1020108 1.1382556 1.8466949
SIG(-3) -2.1029372 1.0998758 -1.9119769
DMR -8.0349240 6.2269175 -1.2903534
DMR(-1) -4.9759800 9.6008864 -0.5182834
-17.303940 10.737224 -1.6115841
-24.481701 9.6090656 -2.5477713
6.1537754 -2.7594848
0.0901606 17.294585
0.0911986 -6.8932053
Mean of dependent
Actual and Fitted Une
A graph of actual, fitted and residual values is part of standard estimation output. You can label graphs in any of 8 fonts and 5 text sizes. MicroTSP allows you to print in 3 sizes in either portrait or landscape orientation.
Percent

MICRO TSP™

Quantitative Micro Software
4521 Campus Drive, #336, Irvine, CA 92715
(714) 856-3368 • FAX (714) 856-2044

THE CHOICE IS CLEAR...

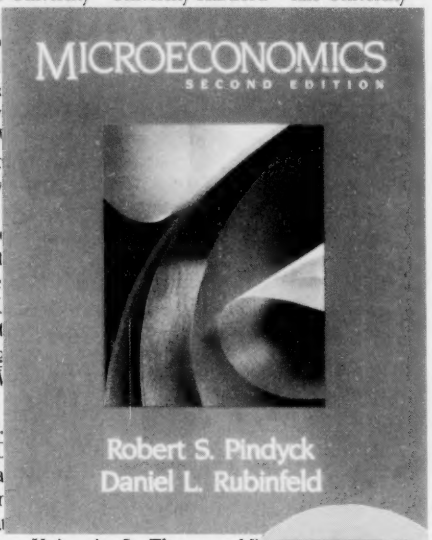
Microeconomics, 2/e **has met the needs of an extremely diverse group of instructors and students across the country. This dynamic and engaging text makes the study of microeconomics pertinent and exciting for students.**

Take a look... see for yourself. This teaching tool will help you unlock your students' potential.

For more information, please contact your local Macmillan representative or write to:

David Borkowsky
Macmillan Publishing Company
866 Third Avenue
New York, NY 10022

ALABAMA: University Alabama Auburn University • ARKANSAS University Arkansas • CALIFORNIA: UC-Berkeley UC-Davis California State University-Fullerton California State University-Long Beach Pepperdine University UCLA University Southern California California State University-Sacramento San Diego State University San Francisco State University University Santa Clara University California-Santa Cruz Stanford University California Lutheran College • COLORADO: University Colorado Metropolitan State University • CONNECTICUT: Western Connecticut State Fairfield University Trinity College Southern Connecticut State University University Hartford Yale University • DISTRICT OF COLUMBIA: Howard University Catho FLORIDA: University Miami Central Florida Florida A& University • HAWAII: Univ University Iowa Buena ILLINOIS: Aurora Universit Illinois-Chicago University University Knox College Illinois State University Bra INDIANA: Indiana Universit University • KANSAS: Unive University Murray State U Tulane University • MASSA University Simmons Colleg University Smith College College Williams College Johns Hopkins UMD St. University Maine Colby C UMI-Ann Arbor UMI-Dea Oakland University Sagir MINNESOTA: Mankato Sta University Bethel College University St. Thomas • M Missouri Northeast Missouri State University Washing University • MONTANA: University Montana • NO Appalachian State University Campbell University D State University Carolina State University Wake F DAKOTA: University North Dakota • NEW HAMP Hampshire • NEW JERSEY: St. Peter's College Rider C Trenton State College • NEW MEXICO: University N University Nevada-Vegas • NEW YORK: SUNY-Albany Lehman College Manhattan College Queens College Fredonia Colgate University Columbia University NYU Stern School of Business SUNY-Oneonta Pace University RIT University Rochester SUNY-Stony Brook Syracuse University Rensselaer PolyTech • OHIO: Bowling Green State University University Cincinnati Xavier University College of Wooster • OKLAHOMA: University Central Oklahoma University Oklahoma Oklahoma City University • OREGON: Oregon State University Portland State University • PENNSYLVANIA: Moravian College Franklin & Marshall Drexel University St. Joseph's University Temple University University Pennsylvania University Pittsburgh Albright College Slippery Rock U/PA Penn State-State College • RHODE ISLAND: Brown University • TENNESSEE: Tennessee Tech University Vanderbilt University • TEXAS: University Texas-Arlington University Texas-Austin South Methodist University University North Texas Texas Christian University Rice University • University Houston Sam Houston State University • UTAH: University Utah • VIRGINIA: George Mason University Hampden Sydney College Hollins College



MICROECONOMICS, 2/e
Robert S. Pindyck, ***Massachusetts Institute of Technology***
Daniel L. Rubinfeld, ***University of California-Berkeley***
1992. 718 pp.
Hardbound.
0-02-396898-1

Announcing a New Handbook for 1993.....

Handbook of Natural Resource and Energy Economics Volume III

edited by **A.V. Kneese** *and* **J.L. Sweeney**

Handbooks in Economics, 6

The *Handbook of Natural Resource and Energy Economics* examines the current theory, and samples current application methods for natural resource and energy economics. This third volume of the Handbook deals primarily with non-renewable resources. It analyzes the economics of energy and minerals, and includes chapters on the economics of environmental policy.

The Handbook provides a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. The surveys summarize not only received results but also newer developments from recent journal articles and discussion papers.

Contents: **Some Basic Concepts.** Economics of Depletable Resources: An Introduction *(J. Sweeney)*. The Optimal Use of Exhaustible Resources *(G.M. Heal)*. Intertemporal Consistency Issues in Depletable Resources *(L. Karp, D. Newbery)*. **Analytical Tools.** Buying Energy and Nonfuel-Minerals: Final, Derived, and Speculation Demand *(M. Slade, C. Kolstadt, R. Weiner)*. Mineral Resource Stocks and Information *(D. Harris)*. Strategies for Modeling Exhaustible Resource Supply *(D. Epple, J. Londregan)*. **Applications to Policy and Forecasting Issues.** Natural Resources in an Age of Substitutability *(P. Dasgupta)*. Natural Resource Cartels *(D. Teece, D. Sunding, E. Mosakowski)*. The Economics of Energy Security: Theory, Evidence, and Policy *(M. Toman)*. Natural Resource Use and the Environment *(C. Kolstad, J. Krautkraemer)*. Energy, The Environment, and Economic Growth *(D.W. Jorgenson, P.J. Wilcoxen)*

1993 about 500 pages
Price: US$ 110.00 / Dfl. 250.00
ISBN 0-444-87800-9

The US$ price is valid in the USA/Canada only. In all other countries the Dutch Guilder ((Dfl.) price is definitive.

NORTH-HOLLAND
An Imprint of Elsevier Science Publishers

ELSEVIER SCIENCE PUBLISHERS
Attn: Heleen A.C. Gierveld
P.O. Box 1991
1000 BZ Amsterdam
The Netherlands

in the USA/Canada:
Attn: Judy Weislogel
P.O. Box 945
Madison Square Station
New York, NY 10160-0757

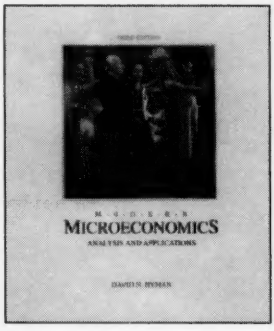
419BNKNEESE2

"I believe the key to understanding the theory is to use it."

--David N. Hyman, regarding his textbook
MODERN MICROECONOMICS: Analysis and Applications

Available Now!

MODERN MICROECONOMICS
Analysis and Applications
Third Edition
David N. Hyman



When we say MODERN, we mean:

- ***New!*** **Modern Microeconomics in Action**--self-contained short essays entitled International Report, The Business World, or Government and the Economy appear in every chapter. These essays provide up-to-date insights, information, and analysis of microeconomic issues.
- **Three New Chapters** covering modern topics:
  Chapter 6: Choices Under Uncertainty
  Chapter 13: Business Strategies in Oligopolistic Markets
  Chapter 17: Asymmetric information and Market Performance
- Coverage of many **new "hot topics"** including: Environmental Protection Policy, TQM, Advances in Technology, New Antitrust Cases and much more!

"If the scale is from 0-3, Hyman's revised text deserves a three plus...Hyman has done a truly outstanding job in developing a readable and applicable text..."

--Dr. Charles G. Leathers, The University of Alabama

IRWIN

EXCELLENCE IN ECONOMICS.

WE SUPPLY WHAT YOU AND YOUR STUDENTS DEMAND.

RYAN C. AMACHER, *University of Texas at Arlington* ◆ **ALAN J. AUERBACH,** *University of Pennsylvania* ◆ **SCOTT BARRETT,** *London School of Business* ◆ **ROBERT A. BECKER,** *Indiana University* ◆ **THEODORE BOS,** *University of Alabama at Birmingham* ◆ **ELBERT V. BOWDEN,** *Appalachian State University* ◆ **WILLIAM J. BOYES,** *Arizona State University* ◆ **GARDNER BROWN, JR.,** *University of Washington* ◆ **WILLIAM F. CHAPELL,** *University of Mississippi* ◆ **REX L. COTTLE,** *Wichita State University* ◆ **DENNIS F. ELLIS,** *University of Michigan, Flint* ◆ **A. MYRICK FREEMAN,** *Bowdoin College* ◆ **DAVID FRIEDMAN,** *Cornell University* ◆ **THOMAS J. HAILSTONES,** *University of South Florida* ◆ **W. MICHAEL HANNEMAN,** *University of California, Berkeley* ◆ **FORREST HARLOW,** *Angelo State University* ◆ **DAVID R. KAMERSCHEN,** *The University of Georgia* ◆ **MWANGI S. KIMENYI,** *University of Connecticut, Storrs* ◆ **MEIR KOHN,** *Dartmouth College* ◆ **LAURENCE J. KOTLIKOFF,** *Boston University* ◆ **FRANK V. MASTRIANNA,** *Slippery Rock University of Pennsylvania* ◆ **KENNETH E. MCCONNELL,** *University of Maryland* ◆ **WILLIAM A. MCEACHERN,** *University of Connecticut, Storrs* ◆ **PAUL NEWBOLD,** *University of Illinois, Champaign-Urbana* ◆ **PAUL R. PORTNEY,** *Resources For the Future* ◆ **FRANKLIN R. ROOT,** *The Wharton School, University of Pennsylvania* ◆ **MARTIN C. SCHNITZER,** *Virginia Polytechnic Institute and State University* ◆ **WILLIAM F. SHUGHART II,** *University of Mississippi* ◆ **GENE SMILEY,** *Marquette University* ◆ **V. KERRY SMITH,** *North Carolina State University* ◆ **DALE B. TRUETT,** *The University of Texas at San Antonio* ◆ **LILA J. TRUETT,** *The University of Texas at San Antonio* ◆ **HOLLEY ULBRICH,** *Clemson University* ◆ **LLOYD M. VALENTINE,** *University of Cincinnati*

TO LEARN MORE ABOUT, CONTACT YOUR LOCAL REPRESENTATIVE OR WRITE:

COLLEGE DIVISION South-Western Publishing Co.

5101 Madison Road
Cincinnati, Ohio 45227-1490

Addison-Wesley and the Santa Fe Institute announce the publication of the first book on the key topic for all finance economists...

The Double Auction Market
Institutions, Theories, and Evidence

Daniel Friedman, University of California at Santa Cruz
John Rust, University of Wisconsin

"With computer networks loosening the grip of traditional intermediary cartels on market structure, the Friedman-Rust compendium takes on considerable significance. Academic theories tested heretofore only in the laboratory are no longer merely academic; one or more of the models analyzed in this book will undoubtedly serve as the basis for successful real markets in the future."

***-Steve Wunsch*, President and CEO, AZX Inc.**

"Popular images of frenzied bidding in double-auction markets obscure their smooth operation in aggregating dispersed information and allocating trades efficiently. The task appears complex but experiments confirm that a few competing traders, each with limited information and computational ability, can approximate a perfect market. Unraveling how markets really work nowadays involves the implementation of competing designs for computerized markets. This volume's comprehensive up-to-date summary of what's known sets an agenda for practical studies of market organization."

***-Robert B. Wilson*, Stanford University**

With seminal papers by:

Colin Camerer
Dan Friedman
John Kagel
John Ledyard
Charles Plott
John Rust
Vernon Smith
Robert Wilson

The Double Auction: Institutions, Theories, and Evidence is available wherever fine professional books are sold or by calling Addison-Wesley at 1-800-447-2226.

Santa Fe Institute

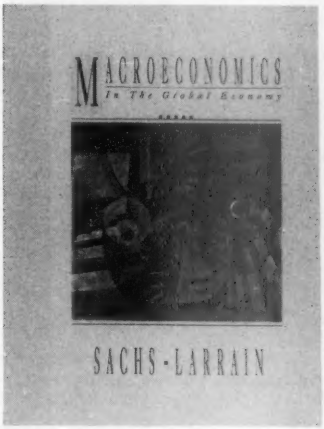
ISBN 0-201-62263-7 Hardcover 419 pages $47.50
ISBN 0-201-62459-1 Paperback 419 pages $24.75

A TRULY OPEN ECONOMY MACROECONOMICS TEXT.

SACHS/LARRAIN

MACROECONOMICS IN THE GLOBAL ECONOMY

Jeffrey D. Sachs, Harvard University, and Felipe B. Larrain, Pontifica Universidad Catolica de Chile



MACROECONOMIC THEORY IN ACTION.
IN POLAND. IN RUSSIA.
IN BOLIVIA. IN THE WORLD.
TAKE A LOOK.

ECONOMICS:
PRENTICE HALL

For an examination copy, please contact your local Prentice Hall representative, or write to Prentice Hall, College Operations, Englewood Cliffs, NJ 07632.

NANYANG TECHNOLOGICAL UNIVERSITY

SINGAPORE

SCHOOL OF ACCOUNTANCY AND BUSINESS

Faculty Appointments in Economics and Statistics

The Nanyang Technological University is a comprehensive university which offers courses at degree and postgraduate levels. It has 33 academic divisions organised within the present 10 schools. There are some 864 faculty and research staff and the current student enrolment stands at 10,890, including 490 postgraduate students.

The School of Accountancy and Business offers undergraduate degrees in accountancy, and business, graduate degrees by research, as well as MBA degrees. Among its undergraduate business programmes is a programme specializing in Applied Economics.

Applications are invited for faculty positions in Economics and Statistics in the School. Applicants are expected to possess a relevant higher degree, preferably at doctoral level, with expertise in the following areas of specialization:

Economics

- Economic Theory (Micro and Macro)
- Mathematical Economics
- Econometrics
- Money and Banking/Monetary Economics
- Public Finance
- International Economics
- Social Accounting and Economic Development
- Labour Economics
- Industrial Economics
- Financial Economics
- Business Forecasting
- Urban and Transport Economics

Statistics

- Stochastic Processes
- Advanced Quantitative Analysis

Sound teaching, professional and research experience are also criteria in recruitment. The successful candidate is also expected to undertake curriculum development work, actively worked at establishing links with the industry within and outside Singapore, and implement management development programmes.

Gross annual emoluments range as follows:

Professor	**: S$108,870 - S$146,970**
Associate Professor	**: S$ 88,650 - S$122,870**
Senior Lecturer	**: S$ 58,680 - S$100,310**
Lecturer	**: S$ 50,390 - S$ 64,200**

(US$1 - S$1.62 approximately)

The commencing salary will depend on the candidate's qualifications, experience and the level of appointment offered.

In addition to the gross annual emoluments, the University adopts the Government's practice in the payment of annual variable component/allowance, the quantum of which is tied to national economic performance and has, in past years, ranged from 1 to 3 months' salary. Other benefits, depending on the type of contract offered, include provident fund benefits or an end-of-contract gratuity, settling-in allowance, subsidised housing, children's education allowance, passage assistance, and baggage allowance for transportation of personal effects to Singapore. Leave and medical benefits will also be provided. Staff members may undertake consultation work of a specialist nature, subject to the approval of the University, and retain consultation fees up to a maximum of 60% of their gross annual emoluments in a calender year.

Applicants should send their curriculum vitae, including the names and addresses of three referees to:

Director of Personnel
NANYANG TECHNOLOGICAL UNIVERSITY
Nanyang Avenue
Singapore 2263
Telefax: (65) 7919340
Internet: Kelly@admin.ntu.ac. sg

PRENTICE HALL ECONOMICS 1993

Thompson/Formby
Economics of the Firm, Sixth Edition
Arthur A. Thompson Jr. and John P. Formby, both of The University of Alabama

Bails/Peppers
Business Fluctuations: Forecasting Techniques and Applications, Second Edition
Dale G. Bails, Memphis State University, and Larry C. Peppers, Washington and Lee University

Just/Hueth/Schmitz
Economics and Public Policy, Second Edition
Richard E. Just and Darrell L. Hueth, both of the University of Maryland, and Andrew Schmitz, University of California—Berkeley

Duetsch
Industry Studies
Larry L. Duetsch, University of Wisconsin—Parkside

Heilbroner
The Making of Economic Society, Ninth Edition
Robert L. Heilbroner

Blau/Ferber
Economics of Women, Men and Work, Second Edition
Francine D. Blau and Marianne A. Ferber, both of University of Illinois at Urbana-Champaign

ECONOMICS:
PRENTICE HALL

For an examination copy, please contact your local Prentice Hall representative, or write to Prentice Hall, College Operations, Englewood Cliffs, NJ 07632.

Oxford

Understanding Consumption

ANGUS DEATON

"The book does an impressive job of integrating the insights available from many sources, drawing on theory as well as econometric evidence, and micro as well as macro models. *Understanding Consumption* is lucid, intuitive, and eminently readable."—Marjorie Flavin, *University of Virginia.*

(Clarendon Lectures in Economics)
1992 256 pp. paper $18.95 cloth $35.00

Understanding the Market Economy

ARNE JON ISACHSEN, CARL B. HAMILTON and THORVALDUR GYLFASON

This book offers a comprehensive analysis of the workings of a planned economy versus that of a market economy. It expounds the principles of market economics in the context of the main problems facing economies in transition from a centrally planned to a market structure.
1992 264 pp.; illus. paper $18.95 cloth $45.00

New in paperback!

Dominating Knowledge

Development, Culture, and Resistance

Edited by FRÉDÉRIQUE APFFEL MARGLIN and STEPHEN A. MARGLIN

These essays examine the role of Western technology in development, and its effect on traditional cultural values and social practices.

(WIDER Studies in Development Economics)
1990 (paper 1993) 312 pp.
paper $24.95 cloth $49.95

Second Thoughts

Myths and Morals of US Economic History

Edited by DONALD N. McCLOSKEY

"The caliber of the reasoning and the presentation of the historical evidence will place this book at the center of economic policy debates. It will be read by those who agree with the argument, and more important, by those who expect to disagree with it. *Second Thoughts* is an excellent volume to be used by students as well as by general readers."—Stanley R. Engerman, *University of Rochester*; coauthor of *Time on the Cross.*
1993 224 pp. $24.95

New in paperback!

Theorists of Economic Growth from David Hume to the Present

With a Perspective on the Next Century

W. W. ROSTOW

"This is the rich and illuminating work we expect from the pen of Professor Rostow, as one of the persons most knowledgeable about both economic growth and the history of economic ideas. No one who works in these fields can afford to overlook this book."—William J. Baumol, *Director, C.V. Starr Center for Applied Economics.*
1990 (paper 1992) 736 pp.; illus.
paper $19.95 cloth $49.95

Fatal Tradeoffs

Public and Private Responsibilities for Risk

W. KIP VISCUSI

"This is an important contribution to our understanding of the economics and social policy of controlling the risks to life. Viscusi has done outstanding work in this field, and much of it is represented here. The treatment is insightful and comprehensive. *Fatal Tradeoffs* is essential reading for anyone who wants the most current information and assessment of this important subject."
—Sherwin Rosen, *University of Chicago.*
1992 320 pp.; illus. $37.00

Prices are subject to change and apply only in the U.S.
To order, send check or money order to: Social Sciences Marketing, Dept. PK.
To order by phone using major credit cards please call 1-800-451-7556.

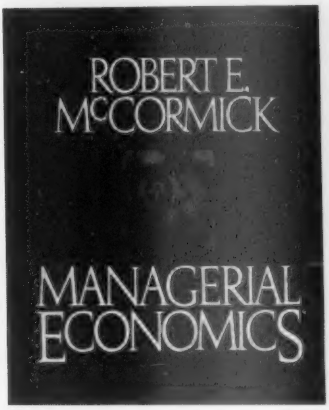
Oxford University Press
200 Madison Avenue • New York, NY 10016

ECONOMICS, COORDINATION AND CONTROL

McCORMICK

MANAGERIAL ECONOMICS

Robert E. McCormick, Clemson University



PROPERTY RIGHTS. INFORMATION COSTS.
VERTICAL INTEGRATION. LOTS MORE.
MODERN STUFF. TAKE A LOOK.

ECONOMICS:
PRENTICE HALL

For an examination copy, please contact your local Prentice Hall representative, or write to Prentice Hall, College Operations, Englewood Cliffs, NJ 07632.

Economically Speaking

Making a Market
The Institutional Transformation of an African Society
Jean Ensminger

In this book, Jean Ensminger analyzes the process by which the market was introduced into the economy of a group of Kenyan pastoralists. She employs neoinstitutional economic analysis in order to assess the impact of new market institutions on production and distribution, with particular emphasis on the effect of institutions on decreasing transaction costs over time. She is able to trace the effects of increasing commercialization on the economic well-being of individual households, rich and poor alike and to analyze the process by which institutions themselves are transformed as a market economy develops.

Political Economy of Institutions and Decisions

42060-1 Hardcover $49.95

The Methodology of Economics
Or How Economists Explain
Second Edition
Mark Blaug

"This is a well-written and stimulating book... The summaries of the relevant literature are very good; its critical comments are balanced and fair. Many economists will find it a useful book." ***—Journal of Economic Literature***

"...If one is looking for a single book that might safely be put into the hands of economics students to inform them about relevant developments in the philosophy of science, about the tradition of self-conscious appraisal in their own discipline, and about the empirical substance of what they will almost everywhere be taught as mainstream economics, then this is it."

—Neil De Marchi, The Economic Journal

43061-5 Hardcover $54.95 / 43678-8 Paper $17.95

Institutions and Social Conflict
Jack Knight

Many of the fundamental questions in social science entail an examination of the role played by social institutions. Why do we have so many social institutions? Why do they take one form in one society and quite different ones in others? In what ways do these institutions originally develop? And when and why do they change? This book addresses these questions.

Political Economy of Institutions and Decisions

42052-0 Hardcover $54.95 / 42189-6 Paper $16.95

Economic Transformation in Eastern Europe and the Distribution of Income
A. B. Atkinson and M. Micklewright

Who gains and who loses from economic transformation in Eastern Europe is a key question—but one that is too rarely understood. This book assembles evidence about earnings, dispersion, income inequality and poverty in Czechoslovakia, Hungary, Poland, The USSR, Russia, The Ukraine, etc.

43329-0 Hardcovr $64.65 / 43882-9 Paper $18.95

Theories of Political Economy
James A. Caporaso and David P. Levine

"Caporaso and Levine skillfully map out the terrain of political economy....The consideration of policy issues nicely illustrates and complements the more abstract points. ***Theories of Political Economy*** combines the virtues of both monograph and text and will edify scholars and students of political economy and policy studies." ***—David F. Weiman, Yale University***

This exploration of some of the more important frameworks used for understanding the relationship between politics and economics includes the classical, Marxian, Keynesian, neoclassical, state-centered, power-centered, and justice-centered.

41561-6 Hardcover $49.95 / 42578-6 Paper $16.95

Advances in Economic Theory: Sixth World Congress
Jean-Jacques Laffont, Editor

These books comprise papers examining the latest developments in economic theory given at the Sixth World Congress of the Econometric Society in Barcelona in August 1990. This is the latest in a series of collections that cover the most active fields in economic theory over a five year period. These books give the reader a unique survey of the most recent advances in economic theory.

Volume 1

Contributors: *Ken Burmore, Eric Van Damme, Drew Fudenburg, Eddie Dekel, David Pearce, John Moore, Ariel Rubenstein, Mathias Dewatripont, Thomas Palfrey*

Econometric Society Monographs 20

41666-3 Hardcover $54.95

Volume 2

Contributors: *Lawrence Epstein, Milton Harris, Arthur Raviv, Jean Tirole, Patrick Bolton, Darrell Duffice, David Cass, Michael Woodford, Roger Guesnerie, Lawrence Jones, Andreu Mas-Colell*

Econometric Society Monographs 21

43019-4 Hardcover $54.95

Economically Speaking

Capital and Credit
A New Formulation of
General Equilibrium Theory
Michio Morishima

Professor Morishima provides a much-needed synthesis of growth theory and monetary theory, drawing on the work of Schumpeter, Keynes and the prewar neoclassical economists to formulate a capital-theoretic general equilibrium theory.
41840-2 Hardcover $49.95

Public Choice Analysis in Historical Perspective
Alan Peacock, Editor

Alan Peacock poses the question of whether the history of economic thought is an essential part of the training of scholars concerned with the study of public finance. He contends that the perspective gained by studying the origins of public choice analysis can offer an important stimulus to scientific progress.
43007-0 Hardcover $44.95

The Economics of Earnings
Solomon W. Polachek and W. Stanley Siebert

In examining the wages people earn, the jobs they do, and the effect of government labor market policies on wages and employment, this work synthesizes research findings using a human capital framework that no other similar book offers.
36476-0 Hardcover $59.95 / 36728-X Paper $19.95

Japan's Capitalism: Creative Defeat and Beyond
Shigeto Tsuru
Foreword by John Kenneth Galbraith

Japan's defeat in the Second World War was total, but her subsequent reconstruction was extraordinary. In this book one of Japan's most eminent economists gives a comprehensive account of the recovery process, and a unique interpretation of the postwar Japanese economy, both as political economist, and as government advisor over most of the period. He analyzes the welfare significance of Japan's money-oriented affluence, and the emergence of a distinctive corporate capitalism. His conclusion is that Japan's inspired creative response to defeat has itself led to a new set of intractable problems.
Cambridge Studies in Economic Policies and Institutions
36058-7 Hardcover $27.95

Incomparable Worth
Pay Equity Meets the Market
Steven E. Rhoads

Challenging proponents of equal pay for comparable worth, Steven Rhoads argues that implementation has been plagued by critical and insurmountable problems. Where success has been most frequently touted—in Minnesota, England, and Australia—job evaluation results are often arbitrary and political rather than objective. Rhoads cites cases where workers have been rated higher than their bosses, or where others previously paid equally for doing equal work have been assigned unequal pay. These and other bizarre outcomes have created severe problems for public and private managers. But the most significant costs of comparable worth go far beyond the administrative. By substituting administered wages for market-determined ones, it produces gross inefficiency in labor markets.
44187-0 Hardcover $24.95

Also available by Steven Rhoads

The Economist's View of the World: Government, Markets and Public Policy
30160-2 Hardcover $49.50
31764-9 Paper $17.95

Price, Quality and Trust
Inter-firm Relations in Britain and Japan
Mari Sako

Long-term, close trading relations based on trust have come to be regarded as a key to the management of buyer-supplier relations in industry, while the arm's length and adversarial commercial bargain has fallen out of favor. This book examines these contrasting approaches, and the link between them and corporate performance, through a study of the electronics industry in Britain and Japan, and offers a unique integration of multi-disciplinary theorizing and original comparative empirical work.
Cambridge Studies in Management 18
41386-9 Hardcover $49.95

Economically Speaking

Now in paperback...

Leading Economic Indicators

New Approaches and Forecasting Records

Kajal Lahiri and Geoffrey H. Moore, Editors

In recent years, significant new developments have emerged in the field of business cycles and economic indicators. This volume contains twenty-two articles by international experts who are working with new and innovative approaches.

Contributors: *Kajal Lahiri, Geoffrey Moore, Frank de Leeuw, Salih N. Neftci, James H. Stock, Mark W. Watson, Michael P. Niemira, Roy H. Webb, Arnold Zellner, Chansik Hong, Stephen K. McNees, Hermon O. Stekler, Stephen J. Silver, Edward F. Renshaw, Allan P. Layton, Francis X. Diebold, Glenn D. Rudebusch, Victor Zarnowitz, Howard L. Roth, James M. Boughton, William H. Branson, Susmita Dasgupta, Ernst A. Boehm, Philip A. Klein, Feliks Tamm*

43858-6 Paper $22.95

Uncertainty

A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis

M. Granger Morgan and Max Henrion

"*Uncertainty* is a risk analysis textbook that Granger Morgan and Max Henrion have written to serve as a basic text for students as well as a reference work for practitioners and researchers. *Uncertainty* succeeds both in providing an excellent basis for policy analysis and in drawing upon a variety of case studies that will be of interest to risk researchers."

—Journal of Economic Literature

42744-4 Paper $16.95

The Market and Beyond

Information Technology in Japan

Martin Fransman

"Martin Fransman has written a valuable book. This book should be important to, and studied and heeded by anyone with an interest or stake in the information technology field."

—American Journal of Economics and Sociology

"Fransman's analysis shows just how complex the Japanese system is. He also shows how cautious Japanese firms have been in entering into cooperative research agreements."

—Journal of Economic Issues

43525-0 Paper $19.95

The Econometric Analysis of Transition Data

Tony Lancaster

"The analysis of transition/duration data is the next frontier beyond the now well-charted area of qualitative choice and limited-dependent variables. Lancaster's book is an excellent travel guide to this relatively unknown but beautiful land where time-series analysis and qualitative choice merge. Even if they do not decide to stay, economists and social scientists will return enriched from a trip under Professor Lancaster's guidance."

—Geert Ridder, University of Groningen

Econometric Society Monographs 17

43789-X Paper $17.95

Beyond the Miracle of the Market

The Political Economy of Agrarian Development in Kenya

Robert H. Bates

"The book succeeds in an empirically grounded synthesis of political and economic concepts—a synthesis that will challenge the thinking of academics and practitioners alike."

—American Anthropologist

Emphasizing the political and economic origins of institutions and their role in economic development, this book examines the political economy of Kenya from the Mau Mau rebellion of the 1950s to the tragic droughts of the 1980s.

Political Economy of Institutions and Decisions

43792-X Paper $17.95

Two-Sided Matching

A Study in Game-Theoretic Modeling and Analysis

Alvin E. Roth and Marilda A. Oliveira Sotomayor

Two-sided matching provides a model of search processes such as those between firms and workers in labor markets or between buyers and sellers in auctions. This text provides a comprehensive account of recent results concerning the game-theoretic analysis of two-sided matching.

Econometric Society Monographs 18

43788-1 Paper $17.95

Economically Speaking

The Elements of Accounting: An Introduction

Geoffrey Whittington

Assuming no previous knowledge or experience in accounting, this concise introduction to the content and structure of accounting statements and the double entry system emphasizes fundamental principles rather than detailed practical procedures.

41392-3 Hardcover $49.95 / 42449-6 Paper $17.95

Hungary: An Economy in Transition

István P. Székely and David M. G. Newbery, Editors

The economic transformation of Hungary has reached a critical stage. This book presents some of the local arguments and perceptions informing the current debate, and critical examination of these ideas by an international panel of scholars. Chapters address the privatisation program; reform of financial, tax and legal systems; integration into the international financial and monetary systems; labor markets, unemployment and the social safety net; and the political economy of the current economic transformation.

Contributors: *István P. Székely, David M. G. Newbery, László Csaba, Kálmán Mizsei, László Halpern, Renzo Daviddi, Zsigmond Járai, Péter Mihályi, János Standler, John P. Bonin, Rumen Dobrinsky, Paul Seabright, Julia Kiraly, Eva Varhegyi, István Abel, Lucian Ionescu, Colin Mayer, Gábor Oblath, Werner Riecke, L. Alan Winters, Richard Portes, Tamás Sárközy, Jeno Koltay, Athar Hussain, Janos Kollo, Maria Augusztinovics, David Winter, David Begg, R. E. Rowthorn, Sir Adam Ridley*

44018-1 Hardcover $59.95

Finance and Development: Issues and Experiences

Alberto Giovannini, Editor

This volume from the Centre for Economic Policy Research presents the latest theoretical and empirical research by leading experts on the relationship between financial market structure and economic growth in developed and developing countries. The issues discussed are of key importance to the future of economic integration in Western Europe, and the current economic transformation of Eastern Europe, and the future of lending to high-debt developing countries.

Contributors: *Alberto Giovannini, Paul Krugman, Oren Sussman, Bruce C. Greenwald, Alec Levinson, Joseph E. Stiglitz, Daniel Cohen, Yung Chul Park, Riccardo Faini, Giampaolo Galli, Curzio Giannini, Ronald I. McKinnon, Juan Ramón Cuadrado, Guillermo de le Dehesa, Andrés Precedo, Philippe Aghion, Robin Burgess*

44017-3 Hardcover $54.95

Trade Flows and Trade Policy after 1992

L. Alan Winters, Editor

This volume offers a major re-appraisal of the effect of completing the European Community's single-market programme on trade among EC members states and also on the Community's external trade. Individual chapters assess the effects of 1992 on competition if transport costs and taste differences remain; detailed case-studies show that integration has complete effects on trade flows and prices and stress the importance of devising a satisfactory international trade policy for the completed market; other chapters estimate the elasticities of demand for imports.

Contributors: *L. Alan Winters, Michael Gasiorek, Alasdair Smith, Anthony J. Venables, Jan I. Haaland, Victor D. Norman, Cillian Ryan, Ian Wooton, Gernot Kleper, Larry S. Karp, Jeffrey M. Perloff, Joaquim Oliveira-Martins, Jõel Toujas-Bernate, Paul A. Brenton*

44020-3 Hardcover $54.95

Available in bookstores or from

CAMBRIDGE UNIVERSITY PRESS

40 W. 20th St., N.Y., NY 10011-4211
Call toll-free 800-872-7423
MasterCard/VISA accepted.
Prices subject to change.

The Foremost Economic Literature Database!

Published by the
American Economic Association

Online

Economic Literature Index
File 139 on DIALOG®

In-depth coverage of journal literature: citations from the Journal of Economic Literature since 1969, with abstracts since 1984. Also citations of articles in collective volumes since 1979.

On Disc

EconLit
CD-ROM by SilverPlatter®

Over two decades of citations and abstracts from the Journal of Economic Literature on a single disc! Includes abstracts of journal articles and books, citations of articles in collective volumes, and dissertation titles.

In Print

Index of Economic Articles in Journals and Collective Volumes

Over 100 years of economic literature indexed by publication year: 1886 through 1988 in 30 volumes on your reference shelf! Make sure your set is up-to-date.

For information, write or call:

Journal of Economic Literature, P.O. Box 7320, Pittsburgh, PA
15213-0320 (412) 268-3869

Now available for individuals!

AEA Members Save $50

EconLit

A SilverPlatter® CD-ROM published by The American Economic Association

The established, reliable database for literature searches in economics derived from the *JOURNAL OF ECONOMIC LITERATURE!*

Hardware Requirements

IBM PC or compatible	**Apple Macintosh**
640 of RAM	MacPlus, SE, or II series
10 MB hard disk drive	Two MB RAM
PC DOS or MS DOS	A hard disk drive
ver. 3.1 or higher	System 6.02 or higher

MS DOS CD-ROM Extensions 2.0 or higher

ISSO-compatible CD-ROM drive

ORDER INFORMATION — Please check appropriate boxes:

☐ Current disc (please allow 2-3 weeks for delivery of current disc) or
☐ Next available disc (updated disc is available 2-3 months after JEL issue)
☐ IBM/MS-DOS 5 1/4 or ☐ 3 1/2 or ☐ Apple Macintosh

AEA MEMBERS:
☐ $300* U.S. & Canada ($330 elsewhere)

NON-MEMBERS:
☐ $350* U.S. & Canada ($385 elsewhere)

Enclosed is my check for $_____, payable to the

American Economic Association
(checks must be drawn on a U.S. bank, payable in U.S. dollars)

Charge $_____ to:
☐ MasterCard or ☐ VISA

Account # _____

Exp. Date: _____

_____
(Name as it appears on account)

Signature: _____

Name _____

Address _____

_____

City _____ State/Province _____ Zip _____

Country _____ Telephone ( ) _____

***License restricted to personal use. Also includes SilverPlatter Information Retrieval System software for EconLit. Prepaid orders only. CD may be purchased once a year only. No free trials. Postage included in the price. Library and department rates available on request.**

Mail to: American Economic Association
Journal of Economic Literature - Dept. AE
P.O. Box 7320, Pittsburgh, PA 15213-0320

or Call: PH: 412-268-3869
FAX: 412-268-6810

HECK's ECONOMIC LITERATURE DATABASE

For the Personal Computer

EXPANDED VERSION NOW AVAILABLE! ONLY $495!

That's right! Only $495 per year for a site license for the most user-friendly, comprehensive and lightening fast literature database available. And our site license allows for networking or use on any pc's throughout your campus. Plus, the newest version includes more than 200 leading journals.

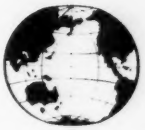
There isn't another literature database anywhere that even comes close. Your institution could easily spend over $2000 per year for the competition. But why? No CD ROM drives needed here! With a few simple commands on your pc, all the relevant references can be retrieved by any keyword.

* Includes over 125,000 bibliographic references from more than 200 leading economic journals from 1950 to present
* Search command compiles listing by author, journal or any keyword contained in an article title
* Narrow command reduces selected list by keyword, giving "and/or" capabilities
* Browse command permits scrolling of index when keyword or its spelling is unknown
* Print/Save command outputs selected references to your printer or an ASCII file
* Typical search time of entire database by keyword is usually less than one second
* Runs on IBM/compatibles with 16 MBytes of disk storage
* Site license allows for use on all campus PCs or networking
* Site license fee only $495 per year with annual updates

To Order, Or For More Information Contact:

Jean Louis Heck
JLH Enterprises
320 West Wayne Ave.
Wayne, PA 19087
Phone/FAX: 215-989-0776

Review of INTERNATIONAL ECONOMICS



RIE

Review of International Economics (RIE) is a new journal devoted to the publication of high-quality articles on topics in international economics: trade theory, commercial policy, economic integration, empirical research (e.g., agricultural trade), international finance, open economy macro issues, trade problems of developing and transition economies, GATT negotiations, etc. Open to controversial and innovative papers. Contributions from related fields such as development and environmental economics are also welcome. Whether theoretical, empirical or policy-oriented, relevance to real world problems is a paramount concern. RIE is produced three times a year by **Blackwell Publishers**.

Contributors (November 1992 issue): Paul Samuelson, Ronald Jones, Ravi Batra, Maurice Obstfeld, Stephen Turnovsky, Marcello Bianconi, and Alex McCalla.

Best Article Awards: RIE will be offering two prizes each year: Best International Economics Article Award ($7,000), and Outstanding International Economics Article Award ($2,000: only assistant professors are eligible).

Associate Editors: Hamid Beladi, Gene Grossman, Paul Krugman, Harvey Lapan; **Book Review Editor**: François Casas. **Editorial Board**: Ravi Batra, Wilfred Ethier, Jeffrey Frankel, Ronald Jones, Giancarlo Gandolfo, Murray Kemp, Stephen Magee, Sugata Marjit, Alex McCalla, Maurice Obstfeld, Michihiro Ohyama, Assaf Razin, Roy Ruffin, Andrew Schmitz, Bruno Solnik, Barbara Spencer, René Stulz, Akira Takayama, Edward Tower, Stephen Turnovsky, Alan Winters.

Manuscripts (**4 copies**) should be sent to: E. Kwan Choi, Editor, Review of International Economics, Dept. of Economics, Iowa State University, Ames, IA 50011. (515) 294-5999; Fax: (515) 294-6336. **Submission fee**: assistant professor, $20; others, $30; none for authors in developing countries).

Subscription Rate: Join the International Economics and Finance Society (IEFS) for a reduced rate. Please send payment ($45 for IEFS membership with subscription to RIE, or $10 for membership only) to **James Hartigan**, Department of Economics, University of Oklahoma, Norman, OK 73019. (405) 325-5501. (For more information about IEFS, please contact **James Hartigan** or **E. Kwan Choi**.)

The general subscription rate is $45 for individuals in N. America and £35 elsewhere. Send payment to: **Paula Stewart**, Journals Marketing, Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF. U.K.

The Experts Agree

"I purchased GAUSS for my PC three years ago. ...After three days exploring the package, I was hooked. I became an avid GAUSS user, and remain so today..."
E.J. Bedrick, Statistics and Computing

"GAUSS is remarkable in its facilities and capabilities. Indeed in the experience of this reviewer the package is unprecedented in its scope, flexibility and power."
Peter J. Dalton, Journal of Economic Surveys

"I love the new debugging facility and the new graphics drivers. I am now using a 1MB Super VGA and it's great to have full resolution available on it. I'm using a 50MHz 486 and the speed of GAUSS is just amazing..."
David Baird, Biometrician, MAF Technology

"The speediest and most versatile of the mathematical programming languages."
"It offers both quick graphics and magnificent presentation-quality graphics."
"GAUSS is a number cruncher par excellence."
Barry Simon, PC Magazine

Maximum Productivity

If you work with numbers, the GAUSS programming language will make you more productive, giving you more answers to more questions, more quickly . . . guaranteed!

Whatever language you are now using, you'll find that GAUSS greatly shortens development time. With GAUSS the basic unit of analysis is a matrix, resulting in a syntax that closely resembles common mathematical expressions. Since matrix operations are assumed, most of the loops required by other languages are not necessary in GAUSS. Over 400 functions including LINPACK and EISPACK routines are provided.

GAUSS for Unix

Now the power of GAUSS is available in the Unix environment. Versions for **IBM RS/6000** and **Sun SPARC** are currently shipping. Call for details and pricing.

GAUSS Light™

If you are looking for an economical yet powerful mathematical tool, GAUSS Light – the fully compatible entry-level version of GAUSS may be just what you need. At just **$150** for DOS and **$295** for Unix, this is an exceptional value.

GAUSS-386i 3.0

Fast and flexible, version 3.0 incorporates many more functions and an intuitive user interface.

Some of the latest features include: Full **32-bit Addressing**, automatic handling of **Complex Numbers**, support for **Virtual Memory** (up to 4 GB), a powerful **Source Level Debugger**, a **Matrix Editor**, flexible **Data Transformation** capabilities, and versatile **Multiple File Viewing/Importing Windows**. GAUSS's 2D and 3D Publication Quality Graphics support SVGA 1024x768 video, and most popular file formats including Color PostScript®.

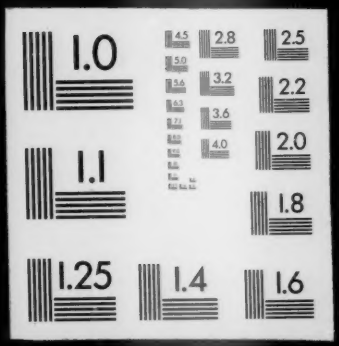
Newly updated **Application Modules** eliminate programming steps for many tasks, while maintaining flexibility for non-standard problems. One of the latest of these is an implementation of the recently introduced Linear Algebra Library – **LAPACK**.

GAUSS-386i 3.0 is just **$995** and comes with a 30-day money back guarantee. Discounts available for Academic use. Call or write for a free brochure on GAUSS Systems and Applications.

GAUSS, GAUSS-386, GAUSS-386i, and GAUSS Light are trademarks of Aptech Systems, Inc. Other trademarks are the property of their respective owners. Prices and specifications subject to change without notice.

Aptech Systems, Inc. Tel: (206) 432-7855, Fax: (206) 432-7832
23804 South East Kent-Kangley Road, Maple Valley, WA 98038 USA

GAUSS™


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8